

VOLUME 46, NUMBER 1

FEBRUARY 1978

JOURNAL OF
CONSULTING
AND
CLINICAL
PSYCHOLOGY

AMERICAN PSYCHOLOGICAL ASSOCIATION

ISSN 0022-006X

The *Journal of Consulting and Clinical Psychology* publishes original contributions on the following topics: (a) development, validity, and use of techniques of diagnosis and treatment in disordered behavior; (b) studies of populations of clinical interest, such as hospital, prison, rehabilitation, geriatric, and similar samples; (c) cross-cultural and demographic studies of interest for the behavior disorders; (d) studies of personality and of its assessment and development where these have a clear bearing on problems of consulting and clinical psychology; (e) case studies pertinent to the preceding topics. In order to improve the use of journal resources it has been agreed by the two Editors concerned that the *Journal of Consulting and Clinical Psychology* will no longer consider manuscripts dealing with the etiology or descriptive pathology of abnormal behavior, and the *Journal of Abnormal Psychology* will no longer consider articles dealing with the diagnosis or treatment of abnormal behavior. Articles that appear to have a significant contribution to both of these broad areas may be sent to either journal for editorial decision. Papers of a theoretical nature will also be considered within the space limitations of the journal.

Brendan A. Maher, Editor
Harvard University
Judith P. Worell, Associate Editor
University of Kentucky

Consulting Editors

- Thomas M. Achenbach, *National Institute of Mental Health, Bethesda, Maryland*
Kent G. Bailey, *Virginia Commonwealth University*
Barbara B. Brown, *Beverly Hills, California*
James N. Butcher, *University of Minnesota*
Jean Chapman, *University of Wisconsin—Madison*
George M. Chertier, *Arizona State University*
Andrew L. Comrey, *University of California, Los Angeles*
W. Grant Dahlstrom, *University of North Carolina, Chapel Hill*
Anthony Davids, *Brown University*
Gerald C. Davison, *State University of New York at Stony Brook*
Alan S. DeWolfe, *Loyola University of Chicago*
Barbara Dohrenwend, *Graduate School and University Center of the City University of New York*
Richard M. Eisler, *Virginia Polytechnic Institute and State University, Blacksburg, Virginia*
Norman S. Endler, *York University, Downsview, Canada*
Sol Garfield, *Washington University in St. Louis*
Donna M. Gelfand, *University of Utah*
Goldine C. Gleser, *University of Cincinnati Medical Center*
Harrison G. Gough, *University of California, Berkeley*
Jesse G. Harris, Jr., *University of Kentucky*
Mary B. Harris, *University of New Mexico*
Kenneth Heller, *Indiana University*
Frances A. Hill, *University of Montana*
Alexandra G. Kaplan, *University of Massachusetts—Amherst*
Alan S. Kaufman, *University of Georgia*
Alan E. Kazdin, *Pennsylvania State University*
Jeffrey A. Kelly, *University of Mississippi Medical Center*
Barbara K. Keogh, *University of California, Los Angeles*
James C. Kinnannon, *University of Minnesota*
Richard I. Lanyon, *Arizona State University*
Herbert M. LeCourt, *University of Waterloo, Waterloo, Canada*
Peter M. Lewinsohn, *University of Oregon*
Michael J. Mahoney, *Pennsylvania State University*
G. Alan Marlatt, *University of Washington*
Albert Marston, *University of Southern California*
Eric J. Mash, *University of Calgary, Calgary, Canada*
Richard McFall, *University of Wisconsin—Madison*
Martha T. S. Mednick, *University of Connecticut*
Barbara G. Melamed, *University of Florida*
David I. Mostofsky, *Boston University*
Benjamin J. Murawski, *Peter Bent Brigham Hospital, Boston, Massachusetts*
Peter E. Nathan, *Rutgers—The State University*
John M. Neale, *State University of New York at Stony Brook*
K. Daniel O'Leary, *State University of New York at Stony Brook*
Oscar A. Parsons, *University of Oklahoma*
Nolan E. Penn, *University of California, San Diego*
N. Dickon Reppucci, *University of Virginia*
Jerome Resnick, *Temple University*
Leonard G. Rorer, *Miami University*
Gerald M. Rosen, *Providence Family Medical Center, Seattle, Washington*
Robert Rosenthal, *Harvard University*
William Schofield, *University Hospitals, Minneapolis, Minnesota*
Carolyn Simmons, *University of Colorado at Denver*
Bonnie Strickland, *University of Massachusetts—Amherst*
Hans H. Strupp, *Vanderbilt University*
Richard M. Suinn, *Colorado State University*
Richard S. Surwit, *Duke University Medical Center*
Patricia B. Sutker, *Medical University of South Carolina*
Norman F. Watt, *University of Massachusetts—Amherst*
Robert L. Weiss, *University of Oregon*
Joan Welkowitz, *New York University*
G. Terence Wilson, *Rutgers—The State University*
Patricia Wisocki, *University of Massachusetts—Amherst*
Melvin Zax, *University of Rochester*

APA Journal Staff

- Anita DeVivo, *Executive Editor*
Kendall C. Patterson, *Production Supervisor*
Deborah S. Felder, *Assistant to the Executive Editor*
Barbara R. Richman, *Production Editor*
Robert J. Hayward, *Advertising Representative*
Juanita Brodie, *Subscription Manager*

Manuscripts: Submit manuscripts in triplicate to the Editor-elect, Sol L. Garfield, Department of Psychology, Washington University, St. Louis, Missouri 63130 according to instructions provided elsewhere in the journal. Authors submitting a manuscript previously considered for publication in another APA journal are invited to inform the Editor, who will then seek to obtain independent reviews, thus avoiding the possibility of repeated reviewing by the same consultant.

Copyright and Permissions: All rights reserved. Written permission must be obtained from the American Psychological Association for copying or reprinting text of more than 500 words, tables, or figures. Permissions are normally granted contingent upon the permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$10 per page, table, or figure of APA material. Permission fees are waived for authors who wish to reproduce their own articles. Address requests for permission to the Permissions Office, American Psychological Association, 1200 Seventeenth Street, N.W., Washington, D.C. 20036.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1971: \$32 domestic, \$34 foreign, \$6 single issue. APA member rate: \$16. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Department, APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P.O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Published bimonthly (beginning in February) in one volume per year by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D.C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D.C., and at additional mailing offices.

347 11/278
Feb

Journal of Consulting and Clinical Psychology

Copyright © 1978 by the American Psychological Association, and *Journal of Coun. & Psychol. Res.*

37
February 1978

Volume 46, Number 1

Articles

- 1 On the Long-Range Effects of Concentration Camp Internment on Nazi Victims.
25 Years Later/*Netta Kohn Dor-Shav*
- 12 Self-disclosure as a Function of State and Trait Anxiety/*Amy L. Post, Bruce C. Wittmaier, and Mitchell E. Radin*
- 20 Effectiveness of Widows' Groups in Facilitating Change/*Carol J. Barrett*
- 32 Reduction of Test Anxiety Through Cognitive Restructuring/*Marvin R. Goldfried, Marsha M. Linehan, and Jean L. Smith*
- 40 Psychological Androgyny and Interpersonal Behavior/*Jerry S. Wiggins and Ana Holzmuller*
- 53 Comparing the Empirical Validity of the Standard Form with Two Abbreviated MMPIs/*Charles S. Newmark, David R. Ziff, Alfred J. Finch, Jr., and Philip C. Kendall*
- 62 Treatment of Female Sexual Dysfunction Through Symbolic Modeling/*Georgia H. Nemetz, Kenneth D. Craig, and Gunther Reith*
- 74 Behavioral Treatment of Problem Drinkers: A Comparative Outcome Study of Three Controlled Drinking Therapies/*William R. Miller*
- 87 Comparison of Usual and Experimental Patients in a Psychiatric Day Center/*Marsha Vannicelli, Stephen Washburn, Betty-Jane Scheff, and Richard Longabaugh*
- 94 Comparative Effectiveness of Day Hospital and Inpatient Psychiatric Treatment/*W. E. Penk, H. L. Charles, and T. A. Van Hoose*
- 102 Test Anxiety and the Passage of Time/*Irwin G. Sarason and Rick Stoops*
- 110 A Cognitive-Behavioral Treatment for Impulsivity: A Group Comparison Study/*Philip C. Kendall and A. J. Finch, Jr.*
- 119 Influence of Contextual Cues on the Efficacy of Desensitization and a Credible Placebo in Alleviating Public Speaking Anxiety/*Jeffrey M. Slutsky and George J. Allen*
- 126 A New Measure of Psychological Androgyny Based on the Personality Research Form/*Juris I. Berzins, Martha A. Welling, and Robert E. Wetter*
- 139 Sensation Seeking in England and America: Cross-cultural, Age, and Sex Comparisons/*Marvin Zuckerman, Sybil Eysenck, and H. J. Eysenck*
- 150 Validation of the Beck Depression Inventory in a University Population Using Psychiatric Estimate as the Criterion/*William Bumberry, J. M. Oliver, and James N. McClure*
- 156 Projective and Repressive Styles of Processing Aversive Information/*Alfred B. Heilbrun, Jr.*

Comments

- 165 The Right to Treatment: Issues in the Treatment of Homosexuality/*Ellie T. Sturgis and Henry E. Adams*
- 170 Not Can but Ought: The Treatment of Homosexuality/*Gerald C. Davison*

(Continued on next page)



Brief Reports

- 173 Modality, Self-disclosure, and Gender as Determinants of Psychotherapeutic Attraction/*Michael R. Kowitt and John P. Garske*
- 175 Factor Analysis of the Ward Atmosphere Scale/*Lynn Alden*
- 177 Defensive Externality and Level of Aspiration/*Dorothy J. Hochreich*
- 179 Predictors of Child Noncompliant Behavior in the Home/*Rex Forehand, Karen C. Wells, and Ellie T. Sturgis*
- 180 MMPI Correlates of Drug Addiction Based on Drug of Choice/*Lee Trevithick and Harmon M. Hosch*
- 181 Predicting Confabulation from the Graham-Kendall Memory-For-Designs Test/*Dan Joslyn, John L. Grundvig, and C. J. Chamberlain*
- 183 Role of Self-instruction and Self-reinforcement in the Modification of Impulsivity/*Wilbur J. Nelson, Jr., and John C. Birkimer*
- 184 The Palmar Sweat Index as a Function of Repression-Sensitization and Fear of Dentistry/*Charles E. Early and Ronald A. Kleinknecht*
- 186 Effects of "Status Sets" on Rotter's Locus of Control Scale/*Kay M. Davidson and Kent G. Bailey*
- 187 Effects of Model Status and Juvenile Offender Type on the Imitation of Self-reward Criteria/*T. John Akamatsu and Parvis A. Farudi*
- 189 Factor Structure and Scale Reliabilities of the Adjective Check List Across Time/*Stephen H. Miller, Charles A. O'Reilly, Karlene H. Roberts, and Carlyle H. Folkins*
- 192 The Seashore Tonal Memory Test as a Neuropsychological Measure/*Carl B. Dodrill and Sureyya S. Dikmen*
- 194 Relationships Between Dimensions of Anxiety and Sensation Seeking/*Barry R. Burkhart, Raymond M. Schwarz, and Samuel B. Green*
- 196 Psychotherapy Process Variables Distinguishing the "Inherently Helpful" Person from the Professional Psychotherapist/*Beverly Gomes-Schwartz and Joseph M. Schwartz*
- 198 "Acceptance," Values, and Therapeutic Change/*Larry E. Beutler, Stephen Pollack, and Avis Jobe*
- 200 Effect of Physical Attractiveness on Therapists' Initial Judgments of a Person's Self-concept/*Steven E. Hobfoll and Louis A. Penner*
- 202 Empathy and Imagery in Avoidance Behavior Reduction/*John T. Esse and Wallace Wilkins*
- 204 Frequent Citations in the *Journal of Consulting and Clinical Psychology* During the 1970s/*W. Miles Cox*
- 206 Predicting Violent Behavior from WAIS Characteristics: A Replication Failure/*Lois Shawver and Charles Jew*

Other

- 11 Instructions to Authors
- 52 Garfield Appointed Editor, 1979-1984

On the Long-Range Effects of Concentration Camp Internment on Nazi Victims: 25 Years Later

Netta Kohn Dor-Shav
Bar-Ilan University, Ramat-Gan, Israel

This study investigated long-range effects of concentration camp internment on survivors. The underlying hypothesis was that the extreme and prolonged stress suffered by victims could be expected to have resulted in impoverishment of personality and dedifferentiation in both personality and perceptual-cognitive functioning. Subjects were 42 survivors, 26 males and 16 females, ranging in age from 42 to 67, and 20 controls, 9 men and 11 women, of similar age, background, and education but who had escaped incarceration. The study was carried out in Israel, using a double-blind paradigm. All subjects were tested with three of Witkin's measures of psychological differentiation, the Embedded-Figures Test, Human-Figure Drawings, and the Block Design subtest of the Wechsler Adult Intelligence Scale. They were also given the Rorschach Inkblot Test, the Bender Gestalt Test, and the 16 Personality Factor Questionnaire. Findings tend to support the hypothesis. Survivors manifested evidence of impoverishment and constriction of personality and appeared to be less accessible, less connected, and more labile. In perceptual-cognitive functioning, they tended to be more global, less complex, and less differentiated, and there were indications of breakdown of ego boundaries. Some evidence suggests that earlier incarceration led to more severe impairment. The price is still being paid!

Although it is more than a quarter of a century since the liberation of Aushwitz in January of 1945, it is still difficult to address oneself objectively to the effects of that holocaust on those few who survived. Nevertheless, it was felt that if the survivors of that ineffable madness were not reached soon, scientific assessment of the sequelae of concentration camp internment would be made impossible by the passage of time.

The present study, therefore, was initiated 25 years after the liberation in an attempt to ascertain the price paid—and still being paid—by the victims of the most inhuman stress experience ever perpetrated by man against man before it became too late to reach the victims.

This study was carried out with the help of funding granted by the Committee for Research of Bar-Ilan University. The author gratefully acknowledges the support given.

Requests for reprints should be sent to Netta Kohn Dor-Shav, Department of Psychology, Bar-Ilan University, Ramat-Gan, Israel.

The psychological literature, of course, reflects considerable interest in the effects of laboratory-induced or situational stress, as well as in the aftereffects of stress experiences such as that of the prisoner-of-war, criminal imprisonment, sensory-deprivation states, and natural and man-made disasters.

However, none of these compare with—and only the prisoner-of-war experience in some cases approached—the degree of stress experienced in the concentration camp, a stress situation that taxed to the extreme human ability to survive in the face of manifold, prolonged, and unprecedented stresses.

It was, of course, recognized immediately after the liberation that the camp experiences had left indelible effects on the few pathetic survivors (Boder, 1949; Grygier, 1954; Nierembski, 1946). Unfortunately, though understandably, however, no scientific, systematic study of survivors has been undertaken over the years. There have been, in fact, only a very few studies, and these tended to focus on attitudes, adjustment, clinical

symptomatology, and life-style—suggesting a unique, long-lasting syndrome among survivors characterized by chronicity, vagueness, depression, fatigue, sleep disorders, emotional lability, blunted affect, and withdrawal (Chodoff, 1963; de Wind, 1968; Niederland, 1968; Simenauer, 1968). Also noted were the absence of psychotic symptomatology and difficulties in interpersonal relationships.

Thus a review of the literature found no studies that attempted to assess the intellectual and perceptual-cognitive functioning of the survivors, nor were there studies aimed at a comprehensive and objective evaluation of personality functioning; no attempts to identify, assess, and define the nature of the psychological deficit or to examine the personality structure and function of survivors in a controlled study seemed to have been made. It was to fill this gap that the present study was undertaken.

Specifically, the study made use of clinical psychological tests and research measures in an attempt to assess the personality functioning, ego boundaries, and perceptual-cognitive processes of concentration camp victims.

Underlying hypotheses were that the severe and prolonged stress endured by survivors may be expected to have resulted in impoverishment of personality, a process of dedifferentiation in the perceptual-cognitive sphere, and primitivization in both personality and perceptual-cognitive functioning.

Thus the research aimed to answer the following basic questions: (a) Is there evidence of primitivization and/or dedifferentiation in perceptual cognitive or personality functioning of survivors? and (b) Are there demonstrable decrements or differences in personality structure or functioning in a group of survivors as compared to controls? In addition, the study was interested in ascertaining if there was evidence of breakdown of ego boundaries, differential effects related to age at the time of incarceration, or actual brain damage.

Method

Subjects

There were a total of 62 subjects of both sexes in the study, ranging in age from 42 to 67, and living

within the greater Tel Aviv area in Israel. All subjects were tested individually in their homes or places of work. Testing sessions lasted from 2 to 3 hours.

Group 1, the main experimental group, consisted of 42 survivors (26 men and 16 women), who had been incarcerated in the most severe of the Nazi concentration camps during World War II. Subjects in this group were taken from a list supplied by Yad Vashem, the Israeli organization instituted devoted to the memory and study of the holocaust with whom all survivors in Israel are registered. Of those included in the study, 13 had been interned at Auschwitz, 5 at Bergen Belsen, 6 in Maidanek, 1 in Treblinka, and 10 in other camps. Ages in this group ranged from 42 to 67.

Group 2, the control group, consisted of 20 subjects (9 men and 11 women) matched to subject in Group 1 for age, academic training, occupation and place of origin. These subjects were chosen via the population registry and were contacted individually. (As there were, unfortunately, an unexpected number of refusals, the *N* for this group is considerably smaller than that for Group 1. The subjects in this group ranged in age from 42 to 65.)

Measures

Each subject was tested with a battery of tests chosen for their appropriateness for measuring perceptual-cognitive and personality functioning.

Three measures of psychological differentiation as defined by Witkin, Dyk, Faterson, Goodenough, and Karp (1962) were included.

1. The Embedded Figures Test (EFT). This test, introduced by Witkin et al. (1962), is considered to be a basic and reliable measure of psychological differentiation. The subject is required to disembed, from within a complex geometrical constellation, a previously presented simpler figure that has been integrated into the new context.

2. The Human Figure Drawing (HFD), as a measure of psychological differentiation, is scored for degree of sophistication and articulation of the body concept.

3. The Block Design subtest of the Wechsler Adult Intelligence Scale (WAIS) is one of the performance subtests considered by Witkin et al. to be a measure of psychological differentiation. It was chosen for the present study because it is also a measure of abstract functioning and is therefore relevant for the assessment of a hypothesis of brain damage.

In addition, the battery included (a) the Rorschach Inkblot Test, a measure of both perceptual-cognitive as well as of personality functioning; (b) the Bender Gestalt Test, a test of psychomotor as well as of perceptual functioning, included primarily to assess the tenability of a hypothesis of brain damage; and (c) the Sixteen Personality Factor Questionnaire (16 PF): This questionnaire, developed by Cattell (1967), yields 16 factors reflecting independent dimensions of personality functioning. It was chosen because it is one of the few reliable per-

validity questionnaires that has been carefully translated into Hebrew and for which norms are in the process of being established in Israel.

Design and Procedure

To mitigate the effects of possible bias, the study was conducted using a double-blind paradigm. Only one assistant (the one in charge of identifying, locating, and contacting subjects, as well as of providing testers with lists of individuals to be tested and of keeping records) knew to which group a particular subject belonged.

Data were gathered by a total of five testers, working two or three at a time. Each tester was carefully trained and supervised by the chief researcher before being allowed to test subjects in the field. Testers did not know the purpose of the study or the nature of the sample groups, nor did they themselves score the tests in the battery. Testers were given lists of subjects to be tested, and they contacted the subjects and tested them individually at the subject's home or place of business. Testing sessions lasted 2-3 hours. Testers were MA students in psychology, with one exception who was in the last year of BA studies.

Scoring was done by a total of three scorers, all of whom were MA candidates in psychology. Again, in accord with the double-blind nature of the study, the scorers remained ignorant of the purpose of the research, as well as of the nature of the sample groups included.

Interrater reliability for scores used was ascertained and was uniformly high, ranging from .42 to .99. (See next section.) However, in the interest of facilitating data processing, one of the scorers whose own intrarater reliability was excellent ($r = .89$) rescored all protocols. It was these scores that were used in the final statistical analyses.

Scoring

Tests in the battery were scored according to the following criteria: (a) for the EFT, the score used was the total time in seconds that the subject took to disembed the 12 figures, with a maximum of 180 seconds assigned for any one figure or for failure (Witkin et al., 1962). (b) The Block Design subtest of the WAIS was assigned its scaled score value. (c) HFDs were scored according to the 9-point version of the Marlen scale (Dershowitz, 1971; Witkin et al., 1962). (d) Rorschach protocols were scored both according to standard clinical methodology as well as for cognitive-differentiation complexity using specially developed criteria described below. (A detailed account of the scoring method and criteria is under preparation for separate publication.)

Briefly, a number of scores, as well as a differentiation index and complexity ratio, were developed to reflect the degree of differentiation, integration, and complexity of the Rorschach responses given by the subject.

Thus *WA* and *WD* are both scores reflecting undifferentiated whole responses, with the former referring to simple, vague, or global gestalts and the latter to simple aggregates of unintegrated or undefined forms (e.g., "bunches of cotton" on Card 7).

WB and *WC*, conversely, are scores reflecting both differentiated and relatively complex percepts, as well as ones having a good deal of integration; the distinction between the two is that *WB* is assigned to responses that seem to progress from the whole to the parts and *WC* for the reverse process. (*WB* responses are relatively rare in any population.)

In addition, two composite scores were developed. The first of these, the differentiation index, reflects not only the quality of the whole responses but also the number of parts or details of the blots that were specifically included in the percepts, as well as the total response pattern. Thus the higher the index, the better the level of articulation and integration.

Specifically, the formula for the differentiation index used is:

$$\frac{[3 (WA)] + [3 (WB) + ND_L] + [3 (WC) + ND_S] + [3 (WD)] + ND + \frac{1}{2} Nd}{R}$$

where *WA*, *WB*, *WC*, and *WD* reflect the qualities described above, *ND_L* and *ND_S* refer to the large details of these differentiated responses, respectively, and *ND* and *Nd* refer to the total number of large detail (*D*) and small detail (*d*) responses given independent of the whole responses. The denominator, *R*, refers to the total number of responses in the protocol.

The complexity ratio reflects the degree of complexity and differentiation of the Rorschach whole responses by contrasting the number of simple, global amorphous responses to the differentiated, complex ones. Thus the formula for the ratio is

$$\frac{WB + WC}{WA + WD}$$

The higher the ratio, the more complex and articulated the responses.

With regard to the reliability and validity of these specially developed measures, the following was found:

Interrater and intrarater reliabilities for *WA* and *WB* were better than .90 ($p < .02$), and for *WB* and *WC* reliabilities ranged from .56 to .78 ($p < .05$). For the differentiation index reliabilities were better than .88 ($p < .02$). (Reliabilities for the complexity ratio, as it is simply a mathematical function of *WA*, *WB*, *WC*, and *WD*, were not computed separately.)

In addition, in conjunction with an earlier study that may be considered a pretest with respect to the present one, test-retest data yielded reliability coefficients for the scores and indices significant at between .05 and .001.

With regard to validity, of course, the problem is more complex. A cogent argument for face and content validity can be made based on the nature of

the procedures and criteria used. The fact that the measures differentiated between groups, according to prediction, in a number of pretests using independent variables such as age and current stress—as well as in the present study (see Results)—is also considered to be in a sense validating.

The 16 PF was scored for its component factors, and these scores were compared across groups.

Results

Addressing ourselves first to the data relevant to the hypothesis that the severe and extreme stress endured by the concentration camp victims would manifest itself in decreased differentiation, survivors were compared to the control group on those measures that were considered relevant for cognitive-complexity differentiation, that is, the EFT, the Block Design subtest of the WAIS, the HFD, and the especially developed Rorschach measures described in the preceding section. Table 1 summarizes these findings and also includes data for the Bender Gestalt test.

Inspection of the data in Table 1 shows that no differences were found between group means on the EFT, BD, or HFD, nor were there significant differences on the Bender.

With regard to the Rorschach, findings were mixed, with support for the hypothesis emerging only from significant differences on the differentiated *WC* measure, as well as for the complexity ratio, though there was also evidence of trends in the expected direction for the *WA* and *WB* scores, as well as for the differentiation index.

It had been noted, however, as inspection of Table 1 shows, that there was a great deal of variability with regard to the EFT, Witkin et al.'s (1962) basic measure of psychological differentiation, especially among the concentration camp group. Upon analysis, this difference yielded an *F* of 2.28, significant at the .05 level—a finding that in itself indicates, at the very least, that the two groups cannot be considered as having come from the same population. Similarly, a significant difference in variability between the groups was found for the differentiation index ($.05 < p < .10$).

This was confirmed in a nonparametric analysis, in which the groups were each divided into *high*, *medium*, and *low* performers.

Table 1
Comparison of all Concentration Camp Survivors and Control Subjects on Perceptual-Cognitive and Differentiation Measures

Measure and group	<i>n</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i> (one-tailed)
Block Design					
CC	40	9.05	2.40		
Control	20	8.75	2.22	.47	.32
Embedded Figures					
CC	35	1,028.57	572.5		
Control	15	1,009.10	399.2	.12	.45
Bender Gestalt					
CC	42	72.26	19.94		
Control	20	66.10	18.34	.79	.22
Differentiation index					
CC	42	2.55	.57		
Control	20	2.70	.78	.83	.21
WA					
CC	42	4.23	2.58		
Control	20	3.40	2.03	1.28	.11
WB					
CC	42	.81	.90		
Control	20	1.00	.85	.77	.22
WC					
CC	42	1.02	1.24		
Control	20	1.55	1.53	1.45	.08*
WD					
CC	42	.60	.95		
Control	20	.70	1.08	.35	.37
Complexity ratio					
CC	42	.50	.53		
Control	20	.84	.80	1.70	.05**
Human Figure Drawings					
Males					
CC	38	7.31	2.29		
Control	14	7.92	1.90	.89	.19
Females					
CC	37	7.35	2.09		
Control	14	7.78	2.08	.66	.26

Note. CC = concentration camp.

* Asterisks indicate significance.

The difference between the groups was found to be significant at the .10 level of confidence (one-tailed).¹

With regard to the HFDs that had not

¹ It should be noted that due to the considerable difficulty, in general, in obtaining significant differences in clinical research, it was decided to accept the 10% level of confidence.

in this analysis, shown any significant differences between groups when scored according to the Marlens scale, it became apparent during subsequent content analyses (not reported here) that the drawings of the concentration camp group tended to be rendered with fragmented broken lines, with vague gestalt, and with inconsistency in the level of detail. Thus, such drawings might include secondary features or items of dress, and even accessories, while omitting essentials. Such differences, however, did not necessarily reflect themselves in the scoring, as a drawing with fragmented lines, which indicated sex and included most body parts, although having obvious evidence of a breakdown of the body boundary concept might well get a better score, than a relatively "healthy" stick figure or a "snowman." These stick figure and snowman styles, however, though obviously undifferentiated and rather primitive, do not reflect a breakdown of ego boundaries. It was decided, therefore, to submit the drawings to a senior clinical psychologist with instructions to separate them into two groups, those that he thought belonged to concentration camp victims and those not. The prediction was clearly significant, $\chi^2(1) = 9.47$, $p < .01$, indicating that on a global comparison, clinical differences are apparent and that the drawings can be differentiated.

At this point—especially due to the variability of the scores—it was decided to ascertain whether age was a factor that should be taken into further account. With an age range of from 42 to 67, it was reasonable to assume that there were a considerable number of subjects included for whom a diminution of powers might be expected on the basis of age alone.

Accordingly, both concentration camp and control subjects were divided into above 50 and 50 and below brackets, and comparisons were made for the scores. This analysis confirmed that age was indeed a factor on the differentiation index, as well as for the Bender.

This led to a replication of all analyses for the perceptual-cognitive and differentiation measures, excluding subjects above 50 in both groups. Table 2 presents the findings.

As will be noted from Table 2, excluding

Table 2
Comparison of Subjects Aged 50 and Below on the Perceptual-Cognitive and Differentiation Measures

Measure and group	<i>n</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i> (one-tailed) ^a
Block Design					
CC	17	8.88	1.79	.48	.32
Control	8	8.50	2.00		
Embedded Figures					
CC	15	1,014.59	621.30	.45	.33
Control	5	877.20	480.65		
Bender Gestalt					
CC	17	70.23	19.75	1.60	.06*
Control	8	57.12	17.75		
Differentiation index					
CC	17	2.36	.47	1.36	.09*
Control	8	2.65	.34		
WA					
CC	17	4.64	3.12	1.12	.14
Control	8	3.37	.19		
WB					
CC	17	.66	.90	.58	.28
Control	8	.87	.64		
WC					
CC	17	1.00	1.02	2.35	.02**
Control	8	2.25	1.66		
WD					
CC	17	.55	1.04	.44	.33
Control	8	.37	.74		
Human Figure Drawings					
Males					
CC	15	6.93	2.31	1.47	.08*
Control	7	8.28	.95		
Females					
CC	15	6.80	2.48	1.53	.07*
Control	7	8.28	.75		

Note. CC = concentration camp.

* Asterisks indicate significance.

the older subjects, as expected, sharpened the findings, resulting in considerable support of the hypothesis. For Witkin et al.'s measures of psychological differentiation it was found that although findings for the EFT were not significant, the Marlens ratings for the HFD yielded significant differences for both the male and female drawings in the predicted direction.

With regard to the special Rorschach scores, the differentiation index yielded a significant difference between the groups, the level of significance for the WC measure was

Table 3
Comparison of all Concentration Camp Survivors and Control Subjects on Selected Rorschach Determinants

Score and group	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i> (one-tailed)*
Productivity				
CC	18.92	11.85		
Control	20.70	11.60	.55	.29
Form %				
CC	69.92	19.48		
Control	69.60	14.26	.07	.48
Form + %				
CC	75.45	11.81		
Control	77.60	7.86	.74	.23
Human Movement				
CC	1.44	1.43		
Control	2.05	1.50	1.53	.07*
Form Primary				
CC	.72	1.12		
Control	.30	.57	1.58	.06*
Color Primary				
CC	.62	.95		
Control	.55	1.23	.27	.39
Pure Color				
CC	.41	.68		
Control	.20	.41	1.36	.09*
Texture				
CC	.46	.59		
Control	.72	.49	1.70	.05**
Sum C				
CC	1.49	1.82		
Control	1.30	1.62	.41	.34
Fisher Penetration				
CC	2.17	2.45		
Control	2.90	2.34	1.14	.13

Note. CC = concentration camp. $n = 42$ for CC subjects, and $n = 20$ for controls.

* Asterisks indicate significance.

increased from .08 to .02. (No computation was made for the complexity ratio for the younger subjects taken alone, since it was apparent that any differences obtained would reflect the difference on *WC* with some contribution from the trend ($p < .20$) for an increase in *WA* in the concentration camp group.)

For the Bender, too, the elimination of the older subjects resulted in a significant difference between the groups, $t(24) = 1.60$, $.05 < p < .10$, one-tailed.

Of considerable interest, of course, was the

analysis of the Rorschach in accordance with the usual scoring methods. Thus the Rorschach protocols were rated not only for cognitive-differentiation complexity but also for the generally accepted formal and content criteria. Tables 3 and 4 present the results for Rorschach determinants and content, respectively, for the concentration camp and control groups as a whole. Table 3, in addition, presents data for the Combined Color (Sum C) and Fisher Penetration scores.

As will be seen from the table, there were significant differences between the groups in the predicted direction for the Human Movement ($.05 < p < .10$), Texture ($p < .05$), and Pure Color ($.05 < p < .10$) determinants, respectively. In addition, there was a significant difference ($.05 < p < .10$) on the Form

Table 4
Comparison of all Concentration Camp Survivors and Control Subjects on Selected Content Aspects of the Rorschach

Content and group	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i> *
Human				
CC	15.21	12.10		
Control	16.76	12.78	.47	.32
Animal				
CC	51.07	17.84		
Control	45.29	16.48	1.25	.12
Anatomy				
CC	6.28	9.06		
Control	5.48	8.35	.34	.37
Fire				
CC	1.67	4.55		
Control	.81	2.79	.80	.22
Abstract				
CC	.40	1.31		
Control	1.86	3.68	2.33	.01***
Sex				
CC	.63	2.85		
Control	2.76	9.67	1.34	.09*
Blood				
CC	1.42	5.35		
Control	.86	2.99	.45	.33
Nature				
CC	2.14	5.70		
Control	5.00	7.37	1.71	.05**

Note. CC = concentration camp. $n = 43$ for CC subjects, and $n = 21$ for controls.

* Asterisks indicate significance.

Primary measure, favoring the concentration camp subjects. (Preliminary indications with regard to Productivity and the Form + %, reflecting deficits in the concentration camp group, were not confirmed, although slight trends remained in the predicted directions.) The differences for the Sum C— and the Fisher Penetration scores were not significant.

With respect to content, as will be noted, concentration camp victims, as expected, gave significantly fewer *abstract* responses ($p < .01$), *nature* responses ($p < .01$), and sex responses ($.05 < p < .10$). In addition, again as would be expected, they tended to give more animal responses ($.10 < p < .20$) and showed a weak trend toward an increase in incidence of *fire* responses.

Next, because age was a serious factor in our analyses for the cognitive-differentiation complexity Rorschach measures, it was decided to also compare several of the traditional Rorschach measures for the younger subjects only. Table 5 summarizes the findings.

As can be noted from Table 5, this analysis again yielded a significant difference for the Pure Color determinant, as well as a significant difference on Color Primary. Findings for the Texture and the Form Primary determinants, however, were weaker than for the entire sample, though strong trends remained ($ps < .10$ and $.13$, respectively).

Finally, scores on the 16 PF were compared for each scale, both for the whole samples as well as for the younger subjects taken separately. Tables 6 and 7 present these data, respectively.

As can be noted from the tables, only the Imaginative scale showed a significant difference between the concentration camp and control groups, with trends apparent also for the Conscientious and Experimenting factors. Taking the younger subjects alone, only the Shrewd and Controlled scales yielded significant differences, with a trend evident for the Self-sufficient factor also. Adopting the standard meanings of the factors for the purpose of interpretation—as well as adopting a considerable degree of caution—we find that as expected, concentration camp victims were rated as more practical, careful, and conventional than the controls (Imaginative). They

Table 5
Comparison of Subjects Aged 50 and Below on the Selected Rorschach Determinants

Score and group	M	SD	t	p (one-tailed) ^a
Productivity				
CC	22.47	16.02		
Control	20.25	7.72	.37	.36
Form %				
CC	74.41	13.97		
Control	73.62	9.87	.14	.45
Form + %				
CC	76.35	9.79		
Control	79.37	5.63	.81	.21
Human Movement				
CC	.83	1.65		
Control	2.00	.75	.81	.21
Form Primary				
CC	.83	1.24		
Control	.25	.46	1.27	.11
Color Primary				
CC	.44	.85		
Control	.00	.00	1.45	.08*
Pure Color				
CC	.66	.84		
Control	.25	.46	1.31	.10*
Texture				
CC	.50	.48		
Control	.75	.53	1.18	.13

Note. CC = concentration camp. $n = 17$ for CC subjects; $n = 8$ for controls.

* Asterisk indicates significance.

tend to be more conservative, staid, and persevering (Conscientious), as well as being more likely to accept established ideas and more tolerant of traditional difficulties (Experimenting). For the younger subjects only, the differences indicated the victims to be more penetrating, more shrewd, more calculating, and more worldly (Shrewd); less disciplined and more likely to follow their own urges and to disregard protocol (Controlled); as well as tending to be more group dependent and more prone to group adherence than the controls (Self-sufficient).

It should be noted that although no predictions had been made with regard to the subscales of the 16 PF before the project was initiated, they were made before the data were analyzed. All findings, with the exception of those for the Controlled scale were

Table 6
*Comparison of all Concentration Camp
 Survivors and Control Subjects on the Sixteen
 Personality Factor Questionnaire Scales*

Scale and group	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i> ^a
Warm-hearted				
CC	11.15	3.39		
Control	11.90	2.42	.68	.25
Emotionally stable				
CC	12.40	3.98		
Control	13.45	4.10	.75	.23
Assertive				
CC	12.28	4.66		
Control	11.63	4.31	.40	.35
Enthusiastic				
CC	10.65	3.12		
Control	10.09	4.03	.48	.32
Conscientious				
CC	14.65	3.44		
Control	13.27	2.73	1.21	.12
Venturesome				
CC	14.59	6.05		
Control	15.00	4.85	.20	.42
Sensitive				
CC	10.75	3.68		
Control	11.36	3.98	.47	.32
Suspicious				
CC	10.50	2.88		
Control	11.81	4.02	.28	.39
Imaginative				
CC	11.68	2.94		
Control	13.18	2.60	1.49	.07*
Shrewd				
CC	11.62	2.99		
Control	11.09	1.37	.51	.29
Worried				
CC	11.71	4.47		
Control	12.09	3.59	.25	.40
Experimenting				
CC	8.43	2.23		
Control	9.45	3.42	.99	.16
Self-sufficient				
CC	11.25	3.48		
Control	12.18	4.09	.73	.24
Controlled				
CC	12.32	2.93		
Control	12.54	2.62	.17	.43
Tense				
CC	14.15	5.46		
Control	12.90	4.27	.69	.25

Note. CC = concentration camp. *n* = 32 for the concentration camp subjects, and *n* = 11 for the controls.

* Asterisk indicates significance.

predicted, though a number of scales for which predictions had been made did not show significant differences.

Discussion

Findings are discussed first in terms of the hypothesis predicting a deficit in cognitive-differentiation complexity for the concentration camp victims, after which the question of personality functioning and possible impoverishment will be addressed.

It was expected that the severe and prolonged stress undergone by the survivors could be expected to have resulted in a process of dedifferentiation and primitivization in the perceptual-cognitive sphere, as well as in personality structure and function. Obviously, it is not possible to demonstrate dedifferentiation or change retroactively; however, data in support of a hypothesis of differences between the groups on the measures allow us to draw inferences with regard to the hypotheses.

On measures of psychological differentiation taken from Witkin et al. (1962), findings were mixed. For the EFT a significant difference was found in variability between the groups, and chi-square was significant at the .10 level, lending some albeit inconclusive support to the hypothesis. The Block Design subtest of the WAIS did not differentiate, and the HFD, rated according to Marlen's scale, did not differentiate between the groups when the entire age range was included. However, when the older subjects were excluded from the analysis, ratings for psychological differentiation of the drawings yielded significant differences between the groups for both the male and female drawings as predicted.

The above findings, in themselves, indicate some support of the underlying hypothesis.

The failure to find differences on Block Design may be explained by what is, in effect, below-par performance within a rather narrow range for both the concentration camp and the control groups. This finding may be a function of the tendency, found in other studies (Dershowitz, 1971), for Jewish subjects at various age levels to do relatively poorly on the performance subtests of the

WAIS, as well as on other measures of psychological differentiation. Thus the fact that group differences were found for the EFT and HFD lends more support to the hypothesis of dedifferentiation here.

It is from the special, original Rorschach measures of cognitive-differentiation complexity that considerable additional support was obtained for the hypothesis, however. Thus, when the groups as a whole were compared to each other, differences were found on the WC measure, indicating a lower incidence of well-articulated and integrated whole responses among the survivors and an overall lower complexity ratio, which reflected the relative frequency of simple, amorphous, non-articulated responses as compared to differentiated, well-articulated, and integrated ones. Further, when the older subjects were excluded from the analysis, we found that the difference on the differentiation index, which had only shown a trend for the groups as a whole, became significant. This measure, because it takes both the qualitative and the quantitative aspects of Rorschach responding into account, is considered to be the most sensitive of the measures for cognitive-differentiation complexity. Thus the finding of a difference at better than the .10 level in the predicted direction lends considerable support to the hypothesis.

Some difficulty is encountered in addressing the question of whether the eventual amount of impairment was greater for victims incarcerated at an earlier age, as has been suggested in the literature regarding survivors (Eitinger, 1964). Obviously, it is the older group of victims who were interned at a later age; however, comparison of older to younger subjects cannot ignore the effects of the aging process itself, which seems to mitigate the differences between the concentration camp and control groups when taken as a whole. The fact that significant differences with age were found within the control group for several of the perceptual-cognitive measures, that is, the EFT, the Bender, the WC, and the HFD, indicates that the involutinal process per se leads to poorer performance of these measures. It becomes all the more interesting, therefore, to note that for the con-

Table 7

Comparison of Subjects Aged 50 and Below on the Sixteen Personality Factor Questionnaire Scales

Scale and group	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i> ^a
Warm-hearted				
CC	11.21	3.51		
Control	12.60	2.07	.82	.21
Emotionally stable				
CC	13.36	4.12		
Control	13.60	5.55	.10	.46
Assertive				
CC	11.28	5.06		
Control	12.40	4.56	.43	.34
Enthusiastic				
CC	11.42	2.95		
Control	11.20	4.97	.12	.45
Conscientious				
CC	14.14	3.71		
Control	15.00	2.23	.48	.32
Venturesome				
CC	14.57	4.41		
Control	14.80	6.05	.09	.46
Sensitive				
CC	11.07	4.34		
Control	10.80	6.05	.11	.45
Suspicious				
CC	10.21	2.25		
Control	10.80	4.38	.39	.35
Imaginative				
CC	13.28	2.58		
Control	13.40	3.28	.08	.47
Shrewd				
CC	13.00	2.51		
Control	10.40	1.51	2.15	.03**
Worried				
CC	10.64	4.43		
Control	10.80	3.11	.07	.47
Experimenting				
CC	9.07	3.43		
Control	11.00	4.47	1.00	.17
Self-sufficient				
CC	11.42	4.27		
Control	13.40	4.66	.87	.19
Controlled				
CC	11.78	2.72		
Control	13.80	2.16	1.49	.08*
Tense				
CC	13.85	5.46		
Control	13.80	4.81	.02	.49

Note. CC = concentration camp. *n* = 14 for CC subjects, *n* = 5 for controls.

* Asterisks indicate significance.

centration camp victims, but not for the controls, there was evidence of a *reversal* on the differentiation index, with younger subjects having significantly lower indices, $t(41) = 2.24$, $p < .02$. As other research has shown a *decrease* with age for this index, the present finding, in effect, lends some support to the hypothesis that incarceration at an earlier age was more damaging.

At this point, the data for the Bender become of considerable interest as well. It can be recalled that when all subjects were taken together, there was no significant difference between the groups. However, when the older subjects were excluded from the samples, a difference between the concentration camp and control groups did obtain. This finding—especially in view of the fact that here, as in the case of the differentiation index, a significant intragroup difference between older and younger subjects was found in the control group but not in the concentration camp sample—can be interpreted as reflecting greater deficit in the younger subjects. Regarding the possibility of actual brain damage among concentration camp victims, the findings remain inconclusive. Though the Bender might suggest damage in the younger subjects, there is no reason to suspect that head trauma as such would be directed more at younger subjects than at older ones. Furthermore the Block Design scores, which are also sensitive to brain damage when it is present, were, if anything, better for the concentration camp group both when the samples were taken as a whole, as well as when only the younger subjects were included, though the difference was of course not significant.

Looking at personality functioning, it was found, as predicted, that there were significant differences between the groups on a number of standard Rorschach scores—in particular on those that measure full, rich, open, and creative functioning. Thus the Rorschach measure that reflects the quality of the inner life, creativity, and perception of the other (the Human Movement response) indicated that concentration camp survivors gave fewer such responses, reflecting a more constricted, impoverished inner life, decreased creativity, and limitation in the perception of the human

"other." The lower incidence of the Texture determinant shows, as predicted, that the survivors had greater difficulty in being accessible to others and in being connected, aware, and open, whereas the increase in Color Form and Color Primary indicate poorer emotional control. In addition, the data indicating significantly fewer *abstract*, *nature*, and *sex* responses also indicate a more restricted, impoverished inner life, whereas the trend to an increase in the number of Animal responses suggests the possibility that for these subjects, the animal has, in a sense, come to replace the human. This interpretation, though of course highly speculative, is given some support by the fact that the animal responses in these protocols are very frequently given in generic, rather than in specific, terms, something that is relatively rare otherwise. Thus we do not find names of specific animals, but rather, "an animal," "a wild animal," "a beast of some sort," "behemoth," and so on, responses that it is felt reflect the subject's experiences with the wild beast unleashed in man and the absence of man's humanity.

From the 16 PF, we found that the survivors, on the whole, tend to be more conservative, careful, conventional, and practical, as well as more staid and persevering—qualities that are consonant with the picture of greater constriction and lessened differentiation gleaned from the Rorschach data. In addition, the suggestion that the survivors tend more to follow their own urges and to be disregarding of accepted protocol fits the finding of an increase in Color Form and Color Primary on the Rorschach. The finding (Self-sufficient) that the survivors tend to be more group dependent and more prone to group adherence supports suggestions to that effect in the literature (Bondy, 1943) and would seem to indicate a basic seeking for safety and affiliation within a group, rather than genuine interaction and belonging.

Taken as a body, then, the data can be seen as providing evidence pointing to impoverishment and dedifferentiation, among concentration camp survivors, in both perceptual-cognitive and personality functioning. The clinical picture is made even more poig-

nant when we consider that the survivors, though tending to seek affiliation and group belongingness overtly, are impaired in the ability to achieve actual connectedness and openness; having been hurt so badly, they would appear to fear close emotional contact. In addition, the breakdown of ego boundaries suggested by the HFD, as well as the dedifferentiation process, can be expected, of course, to make relating objectively difficult.

To summarize: The underlying hypothesis that prolonged and severe stress leads to impoverishment, primitivization, and dedifferentiation seems to have found support in a good deal of the data; the price is still being paid!

References

- Boder, D. P. *I did not interview the dead*. Urbana: University of Illinois Press, 1949.
- Bondy, C. Problems of internment camps. *Journal of Abnormal and Social Psychology*, 1943, 38, 453-475.
- Cattell, R. B. 16 PF. Champaign, Ill.: Institute for Personality and Ability Testing, 1967.
- Chodoff, P. Late effects of the concentration camp syndrome. *Archives of General Psychiatry*, 1963, 8, 323-333.
- Dershowitz, Z. Jewish subcultural patterns and psychological differentiation. *International Journal of Psychology*, 1971, 6, 223-231.
- de Wind, E. The confrontation with death. *International Journal of Psychoanalysis*, 1968, 49, 302-305.
- Eitinger, L. *Concentration camp survivors in Norway and Israel*. London: Allen & Unwin, 1964.
- Grygier, T. *Oppression*. London: Routledge & Kegan Paul, 1954.
- Niederland, W. G. Clinical observations on the survivor syndrome. *International Journal of Psychoanalysis*, 1968, 49, 313-316.
- Nierembski, M. Psychological investigation of a group of internees at Belsen Camp. *Journal of Mental Sciences*, 1946, 92, 60-74.
- Simenauer, E. Late psychic sequelae of man-made disasters. *International Journal of Psychoanalysis*, 1968, 49, 306-309.
- Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R., & Karp, S. A. *Psychological differentiation: Studies of development*. New York: Wiley, 1962.

Received September 15, 1976 ■

INSTRUCTIONS TO AUTHORS

Style of manuscripts. Manuscripts must be prepared in the style described in the *Publishing Manual of the American Psychological Association* (2nd ed.). Instruction on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors should submit manuscripts in triplicate and should keep a copy to guard against loss.

Abstracts. Manuscripts of regular articles must be accompanied by an abstract of 100-175 words. Manuscripts of brief reports must be accompanied by an abstract of 75-100 words. All abstracts must be typed on a separate sheet of paper.

Brief Reports. The *Journal of Consulting and Clinical Psychology* will accept Brief Reports of search studies in clinical psychology. The procedure is intended to permit the publication of soundly designed studies of specialized interest or limited importance which cannot now be accepted as regular articles because of lack of space. Several pages in each issue will be devoted to Brief Reports, published in the order of their receipt without respect to the dates of receipt of the regular articles. Most Brief Reports appear in the first or second issue to go to press following their final acceptance.

An author who wishes to submit a Brief Report:

1. Sends the Brief Report, limited to two printed pages and prepared according to the specifications given below.

2. Agrees not to submit the full report to another journal of general circulation.

Specifications. The Brief Report should give a clear, condensed summary of the procedure of the study and as full an account of the results as space permits.

To ensure that a brief report does not exceed two printed pages, follow these instructions for typing: (a) Set typewriter to a 48-space line. (b) Type text. (c) Count all lines except abstracts, title, and by-line, including acknowledgments. If you have exceeded 185 lines, shorten the material. (d) If your brief report barely exceeds 75 lines (one printed page), try to edit to 75 lines exactly. Reports slightly exceeding one page will be edited for length in order to use space economically.

In Brief Reports, headings, tables, and references are avoided or, if essential, must be counted in the 75 lines. This journal no longer requires an extended report, but, if one is available, each Brief Report must be accompanied by the following footnote:

Requests for reprints and for an extended report of this study should be sent to John Doe (give the author's full name and address).

The footnote should be typed on a separate sheet and not counted in the 75-line quota.

Self-disclosure as a Function of State and Trait Anxiety

Amy L. Post and Bruce C. Wittmaier

Kirkland College

Mitchell E. Radin

Hamilton College

The influence of state and trait anxiety on self-disclosure was investigated. Debilitative and facilitative test-anxious subjects participated in a verbal learning experiment under high and low anxiety conditions. Self-disclosure was elicited by a personal information questionnaire. Subjects' responses to the questionnaire items were assessed for breadth or amount of self-disclosure and depth or intimacy of self-disclosure as well as positive-negative self-evaluation by content analysis of their statements. Results confirmed the hypothesis that individuals experiencing state anxiety disclose less than "normals." These results are discussed in light of various conceptual approaches to anxiety.

Although self-disclosure has been studied in relation to a number of personality, social, and situational variables, the effects of emotional arousal, specifically anxiety, on self-disclosure have yet to be experimentally assessed. As self-disclosure is often elicited in situations such as psychotherapy, in which anxiety may be a predominating factor, understanding the role of anxiety in self-disclosure is of considerable importance. The present study, therefore, attempts to determine the influence of anxiety on self-disclosure.

Self-disclosure has been variously defined as making yourself overt, showing yourself so that others can see you (Jourard, 1964), and the

explicit communication to others of some personal information which the others would be unlikely to acquire unless the person himself disclosed it and which is of such a nature that the individual is not likely to disclose it to everyone who asks for it. (Sermat & Smyth, 1973, p. 332)

Self-disclosure is, thus, the communication of information about one's affects, behaviors, and cognitions with the implication that the material disclosed is either secret, intimate, or emotionally charged.

The concept of self-disclosure is quite complex, since it encompasses both the qualifying of verbalizations and the assessment of the content and direction of verbalizations. Cozby (1973) has proposed that the dimensions of self-disclosure are the breadth or amount of information disclosed, depth or intimacy of this information, and the duration of time spent in disclosure. Two different parameters suggested by Chelune (1975) are the affective manner of presentation of the disclosed material and the flexibility of the disclosure pattern. In addition, other authors have focused on the positive-negative self-evaluative aspect of the content of disclosures (e.g., Sarason & Ganzer, 1962). In the present study the breadth, depth, and positive-negative aspects of individuals' self-disclosure were assessed.

The failure of self-report measures of self-disclosure to predict actual self-disclosure in situations has lead several authors to emphasize the need for behavioral assessment of self-disclosure (e.g., Chaiken & Derlega, 1974). However, there are several issues inherent in the scoring of the content of verbalizations in terms of self-disclosure that present problems. These problems include value judgments as to whether statements about feelings are more disclosing than assessed verbalizations, whether statements in

Requests for reprints should be sent to Bruce C. Wittmaier, who is now at the Lancaster Guidance Clinic, 630 Janet Avenue, Lancaster, Pennsylvania 17601.

the present tense are more disclosing than those in the past or future tenses, and whether negative information about the self is more disclosing than positive information. In addition, it has been found that the level of self-disclosure demanded by questions, modeling, or reciprocity in various situations has a profound effect on the level of self-disclosure elicited (Chaiken & Derlega, 1974; Cozby, 1973; Jourard, 1971). In light of these considerations, we have opted to measure self-disclosure by content analysis of statements made in an experimental situation with a relatively high demand for self-disclosure and to assess self-disclosure on the basis of popular consensus by using a sizable sample of raters for this analysis.

The concept of anxiety, too, is a complex one. Anxiety has been defined as "an unpleasant emotional state or condition which is characterized by subjective feelings of tension, apprehension, and worry, and by activation or arousal of the autonomic nervous system" (Spielberger, 1972, p. 482). Spielberger (1972), in attempting to clarify the concept of anxiety, makes a distinction between state and trait anxiety. He stated that "an anxiety state (A-Trait) is not directly manifested in behavior, but may be inferred from the frequency and the intensity of an individual's elevations in A-State over time" (Spielberger, 1972, p. 482). State anxiety (A-State) is thus defined by the stress level of the situation and the individual's experience of it, whereas trait anxiety (A-Trait) is defined in terms of the individual's propensity to experience state anxiety.

Much of the confusion in the anxiety literature is a result of the failure to examine separately the relationship of A-State and A-Trait with behavior (Spielberger, 1972; Wittmaier, 1974). In the present study the self-disclosure of subjects manifesting high or low trait anxiety was assessed under experimentally created conditions of high or low state anxiety.

Alpert and Haber (1960) have shown that trait measures of anxiety that are geared to specific stress situations (e.g., academic testing) have significantly better predictive validity for both manifest anxiety and performance

in those situations than do general trait measures. The Alpert-Haber Achievement Anxiety Test (AAT) is an A-Trait measure composed of two subscales that discriminate responses to anxiety, which improve performance (Facilitating subscale) from those that interfere with performance (Debilitating subscale). Scores on the Debilitating subscale correlate positively with other measures of A-Trait, whereas scores on the Facilitating subscale correlate negatively and thus reflect low A-Trait (Alpert & Haber, 1960). A useful measure of A-State is the anxiety factor of the Mood Adjective Checklist (MACL; Nowlis, 1965).

Although there have been no direct empirical investigations of the influence of anxiety on self-disclosure, research in this area with subjects characterized by other traits related to anxiety suggests that these subjects tend to disclose less than "normal" subjects. One such trait (need approval, or social desirability) may be thought of as apprehension about negative evaluation, analogous to this type of apprehension in test-anxious subjects (see Sarason, 1975). Anchor, Vojtisek, and Berger (1972) found that psychotic patients who scored extremely high on the Marlowe-Crowne Social Desirability Scale (MCSDS; Crowne & Marlowe, 1964) made a significantly lower proportion of self-statements in a group therapy session than did those scoring low or moderately high in social desirability. Anxiety over negative evaluation appeared to outweigh the tendency to comply with the demand characteristics of therapy for high-need-approval individuals. Similarly, Burhenne and Mirels (1970) found high need for approval to be correlated with low self-disclosure on five essay-type questions. Kopfsstein and Kopfsstein (1973) were unable to find a significant correlation between self-disclosure and social desirability as measured by the MCSDS, but they did find that high-need-approval subjects, as measured by other scales, were more impersonal and more evasive than their low-need-approval counterparts. Subjects with another trait related to general anxiety, uncertainty anxiety, were found to anticipate having a more superficial conversation and more per-

sonal discomfort for a subsequent conversation with an interviewer (Doster & Slaymaker, 1972). Sarason and Ganzer (1962) reported that high-test-anxious subjects presented more negative self-references than low-anxious subjects and responded to evaluative threat with increased negative self-references when asked to describe themselves prior to evaluative threat and no-threat testing.

Other evidence suggests that trait anxiety may influence self-disclosure such that the anxious individual does not modulate level of disclosure to correspond to situational demands. The anxious individual may be characterized by overly consistent patterns of self-disclosure. Chaiken, Derlerga, Bayma, and Shaw (1975) measured the self-disclosure of normal subjects and subjects high in neuroticism in response to a high- or low-disclosing confederate. These authors found a significant interaction between neuroticism and experimental condition such that the disclosure of normal subjects was considerably higher in the high-confederate-disclosure condition than in the low condition, whereas neurotic subjects maintained a moderate level regardless of condition. Finally, since communication involving self-disclosure is central to affiliative behavior, research on anxiety states and affiliation is quite relevant to the study of the effect of state anxiety on self-disclosure. Sarnoff and Zimbardo's (1961) finding that the desire for affiliation decreases with increased anxiety suggests that subjects in a high-anxiety condition will also disclose less about themselves.

It was therefore hypothesized that subjects in this experiment with higher debilitating anxiety, as well as those in a high anxiety state, would disclose less about themselves than other subjects.

Method

General Design

Subjects were identified as debilitators (high A-Trait) or facilitators (low A-Trait) using scores on the AAT. After performing a learning task under conditions designed to create high or low state anxiety, they completed a self-disclosure questionnaire. The effectiveness of the manipulations was checked using scores on the Anxiety factor of the Mood

Adjective Check List. Thus, self-disclosure could be assessed as a function of both A-Trait and A-State.

Subjects

Subjects were 48 freshmen males drawn at random from the class of 1979 at Hamilton College. Subjects were solicited by telephone to participate in a "memory experiment" and were paid for their participation.

Pretesting

The AAT (Alpert & Haber, 1960) was administered to all freshmen during orientation. The subject pool was divided between facilitators and debilitators according to scores on the AAT. Facilitators were those scoring at least 27 on the Facilitative scale ($M = 31.16$) and no more than 24 on the Debilitating scale ($M = 19.75$). Those with intermediate scores were not included in either group. Twenty-four individuals from each group participated in this study and were assigned randomly to experimental conditions. No mention of the pretesting was made 5-6 months later when the experiment was conducted.

Procedure

Each subject participated in the experiment individually. Upon entering the room for the experiment, the subject was told by the male experimenter that the study was designed as an investigation of verbal and nonverbal learning and memory.

All subjects then completed the same learning task. A list of 60 words was presented on a memory drum with a 4-sec exposure per word, and the subjects were asked to memorize these words in any manner they wished. Subjects were tested after each trial until at least 30 words were recalled. Those in the low-anxiety condition were then told that they had performed very well on the task.

At this point the procedure for the high- and low-anxiety conditions began to differ considerably. Both groups completed a digits backward task. Subjects were read increasing numbers of digits until they failed to respond accurately. The same number of digits was then repeated. Two successive failures on any number of digits were considered criterion. Subjects were then administered nine more trials, two at one digit less than criterion, three at the criterion level, and four at one digit more than criterion, in random order, for a net failure effect. Subjects in the low anxiety condition were told that the task was very difficult and not to worry about the score. After completion of the task, they were told that they had performed well. Subjects in the high anxiety condition were told that the task measured nonverbal memory and thinking ability. At the end of the digits backward task, the experimenter sympathetically inquired as to the health of

the subject, ostensibly to ascertain whether the subject had performed as well as he could have. He did not tell high A-State subjects specifically that they had performed poorly.

Subjects in the high anxiety condition were then led into a small adjacent room where a recall test for the memorized words was administered. The room had two one-way mirrors and was equipped with an intercom and a clock with a sweep second hand. The experimenter left the subject in the room after explaining how the clock worked and proceeded to give the rest of the directions over the intercom. The experimenter, in an authoritative voice, told the subject that he would have

exactly 50 seconds, no more and no less, for the next task. Your instructions for the task will be given, and then the task will be described. I will say "ready, set, begin," and then I will start the clock. When the 50 seconds are up, I will say "stop." You will immediately stop writing and raise your pen. Do not complete what you are doing and do not start anything new! Do you understand?

The subject was then instructed to put his name on the paper in front of him, but as he began to write the experimenter said "not yet" to make him aware that he was being watched by the experimenter. The subject was then instructed to turn over the paper and was told: "Now you will have 50 seconds to write down all of the words that appeared in the memory drum list. Ready! Set! Begin!"

After the recall test the subject was asked to complete the MACL (Nowlis, 1965) to assess the effectiveness of this anxiety manipulation. The anxiety scale consisted of seven adjectives (uneasy, tense, nervous, jittery, on edge, fearful, and "clutched up"), each rated on a 4-point scale, yielding scores from a low of 7 to a high of 28.

Subjects in the low anxiety condition remained in the main room after the digits backward task. These subjects were given the following instructions:

The next part of this study requires your complete cooperation and understanding. You are playing an important role in this study, since we will be using your scores as control scores. In other words, your scores will be used as the basis of comparison for other students who will be performing similar tasks but under somewhat different conditions. Try to do as well as you can, since we would like to have a realistic measure of the number of items that a person can complete in a certain amount of time. Please put your name on the top of the paper. OK? You will now have about 50 seconds to write down as many of the words as you can remember from the word list you just learned from the memory drum. Please begin and stop when I ask you to. Are you ready? All right, begin.

All of these instructions were delivered without any authoritative tone of voice. When the 50 sec had

elapsed, the experimenter told the subject to "please stop." These subjects then completed the MACL.

Self-disclosure Questionnaire

After completing the MACL following the recall test, all subjects were asked to complete a self-disclosure questionnaire and another MACL. They were then briefed on the purpose of the experiment. The self-disclosure questionnaire consisted of four questions modeled after questions rated to demand moderate to high levels of intimacy from a 40-item Self-Disclosure Questionnaire (in Jourard, 1971). Each question had a positive and a negative pole (e.g., abilities and weaknesses), so that positive and negative self-evaluation could be examined. Two forms of the questionnaire, with the arrangement of questions and positive and negative aspects of questions reversed, were administered to control for order effects. In addition, the ordering of positive and negative parts of questions was randomized within each form of the questionnaire. Subjects were given about 10 minutes to write responses to the questions. The decision as to when to terminate writing was left to the subject; the subject could write as much or as little as he chose.

The questions were presented to the subject as a possibility for the experimenter to learn significant information about the subject's personality rather than as an investigation of self-disclosure per se. The following is one form of the self-disclosure questionnaire used in this study:

Often psychological research ignores most of the experiences, feelings, and thoughts that make up an individual. We believe this leads to many oversights and oversimplifications. We are, therefore, interested in knowing more about you as a person. Please take about 10 minutes to answer the following questions as honestly as possible. Your answers to these questions will be kept strictly confidential.

1. How do you react to other people's criticism and praise of you? What things about you do people tend to criticize and praise?
2. How satisfied are you with different parts of your body, for example, weight, height, build, hair? Do members of the opposite sex find you sexually desirable?
3. What are your academic abilities and weaknesses? How do they relate realistically to your plans for the future?
4. What do you do when you feel depressed? When you feel anxious? When you feel affectionate? When you feel happy?

Breadth or amount of self-disclosure was assessed in terms of the total number of words written by each subject.

Content Analysis

Depth or intimacy of self-disclosure and positive-negative self-evaluation were assessed by content

analysis of protocols from the self-disclosure questionnaire. Subjects' responses to Question 2, the question demanding the most intimate answer as determined by independent raters, were typed in random order on 6 pages of eight statements each. Five of the six sheets were collated into a leaflet. Twenty such leaflets were prepared with all possible combinations of the six sheets to yield 16 separate ratings of intimacy and self-evaluation for each statement. The leaflet comprised a content analysis questionnaire, which was distributed to 20 upper-class college students, chosen at random who were asked to rate each statement on an intimacy and a positive-negative scale.

The intimacy, or depth of self-disclosure, of each statement was rated on a 9-point Likert-type scale from most intimate to nonintimate according to guidelines explicated in the instructions for the content analysis questionnaire. Similarly, raters were asked to assess the subjects' self-evaluations on a 9-point Likert-type scale from most positive to most negative. The following are the guidelines that raters were instructed to follow in assessing the subjects' statements:

Intimacy Scale

1. Most intimate: The person wrote things about himself that were of an extremely personal, emotional, secret, or embarrassing nature.

3. Very intimate: The person wrote things about himself that were quite personal, emotional, secret, or embarrassing, although perhaps not consistently so.

5. Somewhat intimate: The person wrote some things about himself that were personal, emotional, secret, or embarrassing, but he may have been evasive or defensive in responding.

7. Less intimate: The person generally said little about himself that was of a personal, emotional, secret, or embarrassing nature.

9. Nonintimate: The person said nothing about himself that was of a personal, emotional, secret, or embarrassing nature.

Positive-Negative Scale

1. Most positive: The person is extremely positive about himself. He is very pleased with his body and believes members of the opposite sex find him quite attractive.

3. Positive: Although he may have some reservations, the person tends to be positive about himself and appears satisfied.

5. Neutral or ambivalent: The person's positive and negative remarks about himself seem to balance or counteract each other.

7. Although the person may indicate some strong points, he appears dissatisfied or negative about himself.

9. Most negative: The person is very displeased

by his appearance and believes that members of the opposite sex are not attracted to him at all.

Results

Significant main effects for both state and trait anxiety were found on the analysis of variance of scores on the Anxiety scale of the MACL administered immediately after the recall test. A significant main effect for state anxiety was also found on the analysis of variance of scores on the Anxiety scale of the MACL administered after completion of the self-disclosure questionnaire (Table 1). Thus, after the recall test, subjects in the high anxiety condition and those with high debilitating anxiety reported experiencing moods characterized by terms like tense, nervous, jittery and on edge (i.e., anxiety). Subjects in the high anxiety condition continued to experience more of this type of negative affect after completing the self-disclosure questionnaire.

Significant main effects for both state and

Table 1

Means and Analysis of Variance for the Anxiety Scale of the Mood Adjective Check List

Measure	Anxiety condition	
	High	Low
After recall test		
Debilitating	18.67	14.25
Facilitating	14.08	9.91
After self-disclosure questionnaire		
Debilitating	13.25	10.91
Facilitating	11.08	9.33

Analysis of variance			
Source	df	MS	F
After recall test			
State (A)	1	111.02	4.50*
Trait (B)	1	123.52	5.00*
A × B	1	22.69	.92
Error	44	24.70	
After self-disclosure questionnaire			
State (A)	1	363.00	4.09*
Trait (B)	1	.08	.00
A × B	1	27.00	.31
Error	44	88.66	

* $p < .01$, one-tailed.

trait were also found on the analysis of variance of the number of words written on the self-disclosure questionnaire, the measure of the breadth of self-disclosure. As predicted, subjects in the high anxiety condition and those with greater debilitating anxiety wrote significantly fewer words on the questionnaire (Table 2). Intercorrelations of the number of words written per question were significant for all but one comparison. Eight out of nine of these were significant at or beyond the .001 level, indicating a high degree of consistency in each subject's response to the questionnaire and a high reliability for the measure.

The analysis of variance of the depth of self-disclosure as rated on the intimacy scale of the content analysis questionnaire yielded a significant main effect for state. Subjects in the high anxiety condition were seen to have disclosed less intimately about themselves than their control counterparts. There was no significant main effect for trait on this measure nor any interaction effect (Table 3). A Pearson correlation of .75 ($p < .001$) was

Table 2
Means, Analysis of Variance, and Intercorrelations of the Number of Words Written on the Self-disclosure Questionnaire

Condition	Debilitating	Facilitating
High anxiety	207.42	231.75
Low anxiety	235.75	285.25

Analysis of variance			
Source	df	MS	F
State (A)	1	20,090.08	3.90*
Trait (B)	1	16,354.08	3.17*
A \times B	1	1,900.08	.37
Error	44	5,159.65	

Intercorrelation of words written per question				
Item	2	3	4	Total
1. Question 1	.13	.35**	.49***	.68***
2. Question 2		.50**	.52***	.70***
3. Question 3			.53***	.74***
4. Question 4				.88***

* $p < .05$, one-tailed.

** $p < .01$.

*** $p < .001$.

Table 3
Means and Analysis of Variance for Rated Intimacy from the Content Analysis Questionnaire

Condition	Debilitating	Facilitating
High anxiety	6.33	6.70
Low anxiety	5.66	5.63

Analysis of variance			
Source	df	MS	F
State (A)	1	4.91	3.21*
Trait (B)	1	.05	.03
A \times B	1	.02	.01
Error	44	67.38	

* $p < .05$, one-tailed.

found for the split-half interrater reliability on the intimacy scale.

There was no significant main effect on the positive-negative scale of the content analysis questionnaire (Table 4). The State \times Trait interaction approached but did not attain significance, $F(1, 44) = 2.23$, $p < .13$, two-tailed. The split-half interrater reliability was .93 ($p < .001$). Subjects with high facilitating anxiety tended to express more positive self-evaluations in the low anxiety condition than the control condition, whereas subjects with high debilitating anxiety tended to express more positive self-evaluations in the experimental condition than in the control.

Discussion

The effectiveness of the situational cues of the high anxiety condition to elicit state anxiety was borne out by the difference of scores on the Anxiety scale of the MACL completed after the recall test. Those in the high anxiety condition did, in fact, report more anxiety than did those in the control condition. In both conditions subjects with high debilitating test anxiety reported experiencing more anxiety than facilitators. This result gives credence to Spielberger's contention that trait anxiety, in this case debilitating test anxiety, can be operationalized as the propensity to experience state anxiety with greater frequency or intensity, even in relatively innocuous testing situations like the control con-

Table 4
Means and Analysis of Variance for Rated Self-evaluation from the Content Analysis Questionnaire.

Condition	Debilitating	Facilitating
High anxiety	4.47	4.97
Low anxiety	4.98	4.27

Analysis of variance			
Source	df	MS	F
State (A)	1	.11	.06
Trait (B)	1	.13	.07
A \times B	1	4.42	2.23
Error	44	1.99	

dition. Subjects in the high anxiety condition reported experiencing more anxiety than controls even after the self-disclosure questionnaire was administered, though now there were no differences between debilitators and facilitators. This result may be viewed as supporting the notion of the highly situation-specific nature of the trait, debilitating test anxiety.

As predicted, high state anxiety resulted in less breadth of self-disclosure as well as less intimate disclosure. Spence and Spence (1966) have theorized that anxiety functions to activate dominant responses that have a high degree of habit strength. The lower self-disclosure of subjects in the high anxiety condition thus follows, in that introspective, personal, emotional statements are certainly not the most salient responses in a testing situation. Indeed, the physiological arousal involved in the "fight or flight" response (Selye, 1956) seems antithetical to introspection. Sarason (1975) noted, moreover, that the internal focus of attention of high-test-anxious subjects (A-Trait) leads to failure to attend to external task-relevant cues. In the present study, those in the high anxiety condition, who reported experiencing greater anxiety than controls both before and after completing the self-disclosure questionnaire (and to some extent debilitators in the control condition), may well have been attending to their present arousal state. As a result, perhaps, of this internal focus of attention, they failed to attend to information about them-

selves that was relevant to the self-disclosure questionnaire.

Although debilitators in both conditions did produce less disclosure quantitatively, no difference was seen in the intimacy with which debilitators and facilitators wrote about themselves. As noted above, the debilitators in the control condition experienced lower anxiety than those in the high anxiety condition. Their intermediate level of arousal may have functioned to moderate the number of words they wrote but not their overall intimacy. Working on the self-disclosure questionnaire may have served to moderate the anxiety experienced by debilitators. As intimacy was rated on the basis of responses to the second question (on the alternate form, the third), responding to the preceding question(s) may have served to dissipate anxiety and, consequently, to increase intimacy. Although there has been no empirical demonstration that self-disclosure once elicited functions to ameliorate state anxiety, the finding that self-disclosure is greater in a prolonged stress condition than in a control situation (Altman & Haythorn, 1965) suggests that self-disclosure may serve to reduce anxiety.

The lower intimacy of self-disclosure of subjects in the high anxiety condition may have been, above all, a function of the evaluative threat involved in disclosing intimate information about the self in this particular situation. The anxiety manipulation used in this experiment pivoted on inducing threat to self-esteem by making tasks like the digit span ego involving and creating a failure experience. Although, contrary to expectations, this threat did not manifest itself significantly in the subject's self-evaluations, it may have been crucial in the lower self-disclosure of those in the high anxiety condition. Indeed, low levels of disclosure would seem to function to protect the individual from threat to his most secret or emotionally charged behavior, affect, and thought, especially when this threat is a salient feature of the situation. That debilitators in the control condition did not differ from facilitators in the intimacy of their self-disclosures again suggests that they did not discriminate the self-disclosure questionnaire as a threatening stimu-

lus, whereas the recall test was obviously anxiety evoking for debilitators, again supporting the notion of the situation-specific nature of debilitating test anxiety.

Although this experiment has shown that state anxiety arising from both situational cues (A-State) and the individual's propensity to experience anxiety (A-Trait) results in both lower breadth and depth of self-disclosure, much more empirical work is needed on the relationship of anxiety to self-disclosure. For example, Would the pattern of findings be similar to those of the present study if verbal, rather than written, self-disclosure were used (Anonymous reviewer, Note 1)? Understanding this relationship is crucial to developing effective psychotherapeutic interventions for anxious patients as well as for understanding social maladjustments and their consequences for anxious individuals. Although this experiment did attempt to assess the effect of anxiety on self-disclosure, it did not assess the effect of self-disclosure on anxiety. If self-disclosure does aid in coping with anxiety, as was suggested previously, then inducing anxious subjects to self-disclose in an appropriate situation should lower their level of anxiety. This point, too, needs to be investigated.

Reference Note

1. Anonymous reviewer. Personal communication, May 1977.

References

Alpert, R., & Haber, R. N. Anxiety in achievement situations. *Journal of Abnormal and Social Psychology*, 1960, 61, 207-215.

Altman, I., & Haythorn, W. Interpersonal exchange in isolation. *Sociometry*, 1965, 88, 411-426.

Anchor, K., Vojtisek, J., & Berger, S. Social desirability as a predictor of self-disclosure in groups. *Psychotherapy: Theory, Research, and Practice*, 1972, 9, 262-264.

Burhenne, D., & Mirels, H. Self-disclosure in self-descriptive essays. *Journal of Consulting and Clinical Psychology*, 1970, 35, 409-413.

Chaiken, A., & Derlega, V. *Self-disclosure*. Morristown, N.J.: General Learning Press, 1974.

Chaiken, A., Derlega, V., Bayma, B., & Shaw, J. Neuroticism and disclosing reciprocity. *Journal of Consulting and Clinical Psychology*, 1975, 43, 13-19.

Chelune, G. J. Self-disclosure: An elaboration of its basic dimensions. *Psychological Reports*, 1975, 36, 79-85.

Cozby, P. Self-disclosure: A literature review. *Psychological Bulletin*, 1973, 79, 73-90.

Crowne, D. P., & Marlowe, D. *The approval motive: Studies of evaluative dependence*. New York: Wiley, 1964.

Doster, J., & Slaymaker, J. Need approval, uncertainty anxiety, and expectancies of interview behavior. *Journal of Counseling Psychology*, 1972, 19, 522-528.

Jourard, S. M. *The transparent self*. New York: Van Nostrand, 1964.

Jourard, S. M. *Self-disclosure: An experimental analysis of the transparent self*. New York: Wiley-Interscience, 1971.

Kopfstein, J. H., & Kopfstein, D. Correlates of self-disclosure in college students. *Journal of Consulting and Clinical Psychology*, 1973, 41, 163.

Nowlis, V. Research with the Mood Adjective Check List. In S. Tompkins & C. Izard (Eds.), *Affect, cognition, and personality*. New York: Springer, 1965.

Sarason, I. Test anxiety and the self-disclosing coping model. *Journal of Consulting and Clinical Psychology*, 1975, 48, 148-153.

Sarason, I., & Ganzer, V. Anxiety, reinforcement, and experimental instructions in a free-verbalization situation. *Journal of Abnormal and Social Psychology*, 1962, 65, 300-307.

Sarnoff, I., & Zimbardo, P. Anxiety, fear, and social affiliation. *Journal of Abnormal and Social Psychology*, 1961, 62, 356-363.

Selye, H. *The stress of life*. New York: McGraw-Hill, 1956.

Sermat, V., & Smyth, M. Content analysis of verbal communication in the development of a relationship: Conditions influencing self-disclosure. *Journal of Personality and Social Psychology*, 1973, 26, 332-346.

Spielberger, C. D. Conceptual and methodological issues in anxiety research. In C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (Vol. 2). New York: Academic Press, 1972.

Spence, J. T., & Spence, K. W. The motivational components of manifest anxiety: Drive and drive stimuli. In C. D. Spielberger (Ed.), *Anxiety and behavior*. New York: Academic Press, 1966.

Wittmaier, B. C. Test anxiety, mood and performance. *Journal of Personality and Social Psychology*, 1974, 29, 664-669.

Received July 16, 1976 ■

Effectiveness of Widows' Groups in Facilitating Change

Carol J. Barrett
Wichita State University

Seventy urban widows participated in one of three group treatments for a 7-week period or a waiting list control group. Two therapists each led a self-help group, a "confidant" group, and a women's consciousness-raising group. Personality, attitude, and behavioral measures were obtained at pretest, posttest, and at a 14-week follow-up. At posttest, subjects in *all* conditions had significantly higher self-esteem, experienced a significant increase in intensity of grief, and espoused significantly more negative attitudes toward remarriage. Experimental subjects showed significant improvement in their ratings of future health and became significantly less other-oriented in their attitudes toward women relative to the controls. The therapist variable produced few differences in response to treatment. At follow-up, treatment gains were maintained. Life changes were significantly more positive in the women's consciousness-raising groups, and posttest evaluations of the program by these subjects were significantly higher. All treatments resulted in high rates of contact among participants in the group.

Research has demonstrated amply that widowhood is a stage in the life cycle during which multiple stresses occur. The degree of change in one's life necessary for adjustment to the spouse's death is deemed greater than that required for 42 other major life events (Holmes & Masuda, Note 1). There is strong evidence that widowed persons experience impairment in both mental and physical health relative to married persons of the same age (Barrett, 1977). Psychiatric symptoms (Clayton, 1974; Parkes, 1964), mental illness (Bel-

lin & Hardt, 1958), inception of psychiatric outpatient and inpatient services (Robertson, 1974; Stein & Susser, 1969), and suicide (Bock & Webber, 1972; Cosneck, 1966; Segal, 1969) are all more frequent among the widowed than the married.

Disabling illness (Woolsey, 1952) and hospitalization for medical problems (California Department of Public Health, 1958; Rosenfeld, Katz, & Donabedian, 1957; Rosenfeld, Mott, & Taylor, 1951) are also more frequent among the widowed. Physical symptoms were more prevalent among young widows relative to married women (Maddison & Viola, 1968; Parkes & Brown, 1972) but not among the elderly widowed group (Heyman & Gianturco, 1973). In a landmark study, Kraus and Lilienfeld (1959) documented higher death rates among both black and white widowed persons of both sexes, for all age groups and for all causes of death, when compared to married persons of similar age and socioeconomic status. The discrepancy in death rates between widowed and married persons was largest at the younger levels.

There are convincing arguments that such differences in the level of functioning of same-age married and widowed individuals cannot

This article is based on the author's doctoral dissertation research at the University of Southern California in 1974. A report of the preliminary findings was made at the Gerontological Society Meeting in Miami Beach, Florida, in 1973 (*The Gerontologist*, 13, p. 66).

The author appreciates the comments of Jim Birren during the formulative stage of this research, the suggestions of Al Marston during its execution, and the cooperation of members of the Institute of Lifetime Learning in Long Beach, California, who participated in the pilot study. Lisa Pomeroy served as one of the group therapists, and Elwin Barrett and Sally Daniel assisted in the data analysis.

Requests for reprints should be sent to Carol J. Barrett, Box 34, Department of Psychology, Wichita State University, Wichita, Kansas 67208.

be accounted for totally by predisposing characteristics operative prior to marriage, the effects of a shared unfavorable environment, artifacts in reporting procedures, and the possibility of higher remarriage rates by the healthy widowed (e.g., Parkes, Benjamin, & Fitzgerald, 1969; Stein & Susser, 1969). With an occasional exception (Lowenthal, 1964), widowhood has been shown to increase social isolation (Adams, 1968; Berardo, 1970; Marris, 1958). In Lopata's (1973) extensive study of 301 widows over age 50 in the Chicago area, as well as in other research (Nuckols, 1973), loneliness was indeed perceived as the worst problem. A substantial proportion of widowed persons must assume the additional responsibility of single parenthood. (Widows constitute the largest group of single parents; Schlesinger, 1971.) Financial problems are often severe (Nuckols, 1973; Palmore, Stanley, & Cormier, 1963), and legal problems complicate the resolution of grief that must proceed with little institutionalized support (Gorer, 1965). The available evidence, the stress theory of mental illness (e.g., Simon, 1970) and of physical illness (Holmes & Masuda, Note 1), and the application of social role theory to widowhood (Lopata, 1973) support the hypothesis that the experience of widowhood itself may produce the observed physical and emotional distress.

The importance of this position is highlighted by the size of the population currently and potentially at risk. The 12 million widowed persons in this country already number almost 5% of the total population (U.S. Bureau of the Census, 1970). The projected increase in the size of the older population enhances the probability that the proportion of widowed persons will continue to increase. The probability of widowhood is much higher for women, who usually marry men older than themselves, despite the fact that they outlive men by 7-8 years. Approximately three out of four married women will become widows (Lewis & Berns, 1975). Currently there are more than four times as many widows as widowers. Widowhood is seldom a brief episode at the end of the life span; the average duration of widowhood for widows who do not

remarry and who die from natural causes is 18½ years (Carter & Glick, 1970).

In recognition of the special needs of the widowed, the high-risk nature of this group, and the dearth of responsive programming in traditional mental health agencies (Silverman, 1966), a variety of therapeutic programs have recently been developed by mental health professionals (Abrahams, 1972; Miles & Hays, 1975; Silverman, 1970; Woods, 1973; Antoniak, Note 2; Van Coevering, Note 3; Parkes, Note 4) and by lay organizations (e.g., NAIM, a Catholic-sponsored organization; THEOS (They Help Each Other Spiritually), a nondenominational organization based in Pittsburgh; and American Association of Retired Persons' Widowed-to-Widowed Program). Most utilize widowed paraprofessionals and offer individual help to the recently widowed; others arrange discussion groups. Unfortunately, controlled outcome studies of such innovative programs are nonexistent.

The present study sought to develop and evaluate three different therapeutic group interventions for widowed women and to compare the results of treatment to a waiting list control group. Although the propensity of widowhood is higher among older women, research suggests that stresses may be greater for younger widows (Kraus & Lilienfeld, 1959; Parkes, 1964; Robertson, 1974). Hence, the treatment groups were designed to respond to the needs of widows of all ages. Since many of the social and economic stresses of widowhood persist long after the husband's death, no limit was set on the duration of widowhood for participants in the study.

A basic premise in all three treatment groups was that widowed women would be able to help each other cope with the stresses of their situation (Silverman, 1970). The design of each treatment reflected a different research base or theoretical posture. The possible superiority of self-help strategies in comparison with traditional psychotherapy among groups with pervasive common problems (e.g., Hurvitz, 1970) provided a basis for the self-help groups. The advent of peer counseling methods and the research of Lowenthal and Haven (1968) provided the rationale for the

confidant groups. These investigators have shown that older persons who have a confidant have better subsequent mental health than those without one. Since many widows will have lost a primary confidant in their husbands, the development of a compatible widow confidant might be therapeutic, particularly in light of Lopata's (1973) analysis of the difficulties that widows have in making new friends. The advent of the consciousness raising group in the women's movement provided a model for the widows' *consciousness raising* groups. Although consciousness raising groups enjoy immense popularity and personal accounts of their helpfulness to women abound, little systematic research has accompanied them. Since a number of the stresses of widowhood derive from sex roles operative both in marriage and subsequently, a group experience focusing on sex roles was considered to have potential therapeutic value to widows. Since there were no previous outcome studies to guide the selection of dependent variables relevant to widowed persons' services, a variety of potentially useful measures was included.

Method

Subjects

Two hundred and thirty-nine widows who responded to a news item in the *Los Angeles Times* or one of six Los Angeles area newspapers that briefly described the program were invited to an orientation meeting. (Eighty-one subsequent inquiries were referred to local community mental health centers.) One hundred and twenty-six attended. Of these, 83¹ (66%) participated in the program. Ten out of 28 subjects randomly assigned to the waiting list control group, and 25 out of 86 assigned to treatment did not return. Eight subjects who could not participate immediately due to prior commitments were permitted to join the waiting list group (4 did so), but their data were omitted to prevent confounding of subject characteristics with control group status. Eight subjects in the treatment groups who attended fewer than four of the seven sessions (4 confidant subjects, 2 self-help subjects, and 2 consciousness raising subjects) were dropped from the analysis along with 1 control subject later identified as a divorcee. This left a total of 70 subjects, including 17 controls and 53 assigned to treatment groups.

The subjects ranged in age from 32 to 74, with a mean age of 55.7 ($SD = 9.2$ years). The duration

of widowhood ranged from less than 1 month to 22 years, with a mean of 4 years 9 months and median of 3 years 9 months. Subjects widowed 2 years or less comprised about one third of the sample. Ten subjects had been widowed over 10 years. The majority (87%) had been married only once. About a third still had children at home. Twenty-eight were Jewish; 23, Protestant; 10, Catholic; and 9 indicated some other religious category. Fifty-three percent were employed, 30% were retired or not employed by choice, and 17% were unemployed. More than 70% of the sample had at least some college education, and 16% had a graduate or professional degree. The modal current monthly income from all sources was \$600-\$900; the modal household income prior to widowhood was over \$1,500 a month.

Procedure

Each subject attended one of four orientation meetings, which began with a brief lecture on widowhood. They were told that they would be attending one of three types of discussion group focusing on "specific problems of widowhood" (the self-help groups), "the development of friendships" (the confidant groups), or "the roles of women in society" (the consciousness raising groups). Subjects completed demographic information and the pretest dependent measures and indicated their availability for meeting times (without knowledge of which group would meet at which time). A \$15 check payable at the first group session and to be returned at follow-up was required to reduce subject attrition.²

All subjects who attended orientation (except for those who withdrew) were assigned as randomly as possible to one of three treatment conditions or to a waiting list control group. Several factors prevented complete random assignment. The first was a limitation on the availability of subjects. No subject was assigned to a group that it was impossible for her to attend. To prevent confounding of subject characteristics such as employment status with treatment, one group of each of the treatments met on a weekday, and one met on Saturday.

A second factor was the desirability of pairing confidant group subjects with someone who lives nearby to increase the likelihood of continued relationship.

¹ Inquiries were received from all over Los Angeles County. Phone calls from a number of potential subjects hoping for a neighborhood group suggested that the attendance at orientation and the reduced number of actual participants were in part a reflection of the excessive travel time required in the city.

² Informal feedback suggested that this procedure may have backfired. Some potential subjects who attended orientation but did not subsequently join a group may have been deterred by the fee. Several had no checking account.

tionships posttreatment. To ensure geographic proximity, confidant group assignments were made from among the two most heavily represented of the five toll-free Los Angeles County regions designated by the telephone company. Subjects in one group all resided in the San Fernando Valley; subjects in the other group lived in the central city area. (Additional subjects from these regions were assigned to other treatment conditions.)

The third factor was the desirability of beginning the groups as soon as possible after orientation to reduce the likelihood of attrition. Subjects from the first orientation meetings were assigned to the first therapist or the control group; subjects from the final meetings were assigned to the second therapist or the control group. The final number of subjects in each group appears in Table 1.

All experimental subjects began a 7-week treatment group within 1-2 weeks of orientation. Approximately 14 weeks after the posttest (13-15 weeks), and 6 months after the program announcement, all subjects attended a follow-up session. After completion of data forms, the preliminary results of the study were described. Control subjects participated in a treatment group following the collection of posttest data and hence were omitted from the follow-up analysis.

All groups met for 2 hours once a week for 7 weeks at the University of Southern California. Two (nonwidowed) female doctoral students in clinical psychology each led one of three different groups pretested in a pilot study. The purpose of the self-help groups was to encourage participants to help each other find solutions to the problems of widowhood. These groups were told that widowed persons are the experts on widowhood. The therapist's role was to facilitate discussion; all direct requests for advice were referred back to group members, and reticent members were encouraged to share their experiences. The therapist praised group members when specific suggestions were offered to others. The problems discussed were those initiated by the members themselves and included loneliness, grief, single parenting, reduced financial resources and employment difficulties, decisions about living arrangements, strained relationships with relatives and married friends, new relationships with men, and legal problems.

The purpose of the confidant groups was to facilitate the development of a close friendship between pairs of widows. Subjects were paired for the duration of the group in one of two ways. At the close of the first group discussion,³ subjects could confidentially indicate preferences for pair assignments. Those who did not wish to make this decision were assigned primarily on the basis of matched age and duration of widowhood. The group format consisted of an intimacy training task in pairs, during which the therapist moved from pair to pair as needed, followed by a discussion of the experience by the whole group. The intimacy tasks were arranged in gradually increasing difficulty over the 7-week period; they proceeded from a request to

Table 1
Distribution of Subjects

Group	n
Self-help 1	8
Self-help 2	10
Confidant 1	8
Confidant 2	7
Consciousness raising 1	11
Consciousness raising 2	9
All experimental subjects	53
Waiting list control	17
Total no. subjects	70

find three things the pair had in common to a request to share personal problems and to make an explicit offer of help to one's confidant.

The purpose of the consciousness raising groups was to facilitate the participants' awareness of how their experiences as widows relate to them as women. The group structure was modeled after that developed by the Consciousness Raising Committee of the Los Angeles chapter of the National Organization for Women (Freeman, Note 5). Each group was given a list of possible sex role topics of particular relevance to widows⁴; the group selected the topic to be discussed 1 week in advance. Each session consisted of (a) introductory comments about the topic from the leader's own experience, (b) 5-10 minutes for each member in sequence around the group to express her reactions to the topic without interruption, and (c) open discussion by all participants. "No confrontation" was an enforced rule. Members could ask questions after the individual reactions but were not permitted to make comments until their "turns."

Measures

Eighteen personality, attitude, and behavioral measures were obtained by written self-report; 12 at pretest, posttest, and follow-up; 2 at posttest and follow-up; and 4 at follow-up only (see Table 2). Ten measures, comprised of 6 rating instruments and 4 behavioral indices, were developed by me to assess physical, emotional, and social functioning in widowhood.⁵ These included the frequency of phys-

³ The announced beginning topic in all groups was the question of whether widows or widowers experience greater stress. However, both therapists observed that the actual primary topic was the circumstances of the husband's death, often tearfully recounted. A similar phenomenon occurred in Lopata's interviews (Note 7).

⁴ Available from the author on request. Popular topics included: "Does widowhood oppress women?" and "Are you still a wife?"

⁵ Copies of instruments not previously published are available from the author.

Table 2
Dependent Variables

Pretest

Frequency of physical complaints
5-year health prediction
Intensity of grief
Attitude toward widowhood
Attitude toward remarriage
Social role involvement
Self-esteem
Locus of control
Life satisfaction
Attitudes toward women
Other-orientation
Self-orientation
Radical vs. conservatism

Posttest

Pretest plus
Extent of help from group
Extent of learning in group

Follow-up^a

Pretest plus posttest plus
Quality of life change
Number of group meetings
Number of members contacted
Total number of individual contacts

^a Included experimental subjects only.

cal complaints (summed across six ailments that are common in bereavement), a prediction of one's health in 5 years (poor, fair, good, and excellent were assigned scores of 1-4, respectively), intensity of grief (a sum of 12 7-point rating scales assessing feelings of loneliness, inability to cope, sorrow, guilt, anger, and depression), attitudes toward widowhood and remarriage (2 7-point scales), and social role involvement (sum of ratings representing involvement in eight major social roles). A major behavioral measure was the quality of life change since enrollment in the program; responses to an open-ended follow-up question were rated independently on a 5-point scale from extremely positive to extremely negative by two persons blind with respect to treatment condition. Responses were also content analyzed. Additional behavioral measures permitted an evaluation of the extent of social activity among members of the various groups subsequent to treatment. These were the number of spontaneous group meetings during the follow-up period, the number of group members contacted individually, and the total number of individual contacts during follow-up.

Three personality variables were included: Rosenberg's (1962) self-esteem measure; Rotter's (1966) Internal-External Locus of Control (I-E) Scale; and Neugarten, Havighurst, and Tobin's (1961) Life Satisfaction Index A; as well as three measures of attitudes toward women; a 15-item radical versus

conservatism scale adapted from Spence and Helmreich (1972); and Gump's (1972) other-orientation and self-orientation measures. Other-orientation is a composite of three factors representing the views that identity is derived through traditional roles, that woman's role is submissive, and that home orientation and duty to children should be stressed. Self-orientation is a composite of four factors reflecting a need for individualistic achievement and satisfactions, a sense of autonomy and heightened independence, and the beliefs that the traditional role implies some relinquishing of needs for personal fulfillment and that the family is inadequate to completely fulfill one's needs. At posttest and follow-up, subjects evaluated the degree of help received from the group and the degree to which they learned from the group on 7-point scales.

Results

Pretest

Multivariate and univariate analyses of variance performed on the 12 pretested dependent variables resulted in no significant differences between the four conditions (self-help, confidant, consciousness raising, and control groups) at the start of treatment. Nor was there any difference in the average age or duration of widowhood among subjects in the four conditions.

Posttest

To determine whether treatment groups could be combined across therapists, a multivariate analysis of variance and univariate analyses of variance were performed on the pre-post change scores for the six available widowhood functioning variables and the three personality variables, with therapist (two levels) and treatment (three levels) as independent factors. The only significant therapist effect was for attitude toward remarriage, $F(1, 40) = 6.44$, $p < .05$ (see Table 3), and none of the therapist-treatment interactions was significant. Similarly, tests for therapist and therapist-treatment interaction effects on the attitudes toward women change scores were not significant.

Hence, one-way (four-level) multivariate analyses of variance were performed on the pre-post change scores for the nine personality and widowhood functioning variables, and separately for the three attitudes toward

women variables, on all subjects (including the control group) for whom data were complete. The degrees of freedom in the treatment factor were partitioned to enable planned comparisons to assess (a) whether any significant change occurred in the variables from pretest to posttest, (b) whether the change among experimental subjects differed from the change among controls, and (c) whether there were any between-treatment differences in the change scores.

Both multivariate tests for pre-post change across all subjects were significant, $F(9, 51) = 3.09$, $p < .005$, $F(3, 60) = 4.20$, $p < .01$. Subsequent univariate tests revealed a highly significant increase in self-esteem, $F(1, 59) = 14.07$, $p < .001$, a significant increase in the intensity of grief, $F(1, 59) = 9.92$, $p = .003$, significantly more negative attitudes toward remarriage, $F(1, 59) = 8.07$, $p = .006$, a significant reduction in other-oriented attitudes toward women, $F(1, 62) = 7.50$, $p = .008$, and an almost-significant increase in self-orientation ($p = .07$). The mean changes in these variables by treatment appear in Table 3.

Although neither multivariate test of experimental-control group differences was significant, two univariate tests were significant. Experimental subjects showed a more positive change in their 5-year health predictions, $F(1, 59) = 4.12$, $p < .05$, and became less other-oriented than controls, $F(1, 62) = 3.86$, $p = .05$. Treatment means for these variables appear in Table 3. Scrutiny of the group means revealed that the significant reduction in other-orientation across all subjects was

totally accounted for by the substantial change among experimental subjects. There were no significant between-treatment differences on any of the 12 pre-post variables.

A 3×2 multivariate analysis of variance on the two posttest experimental group evaluation measures yielded a significant main effect for treatment, $F(4, 92) = 2.62$, $p < .05$, and no therapist effect or Therapist \times Treatment interaction. Both extent of help from the group, $F(2, 47) = 4.05$, $p < .05$, and extent of learning in the group, $F(2, 47) = 5.03$, $p < .01$, varied with the treatment. Table 4 presents the mean evaluation scores by treatment. The highest ratings on both variables occurred in the consciousness raising groups and the lowest in the self-help groups.

Follow-up

A one-way (three-level) multivariate analysis of variance and univariate analyses of variance on the change scores from pretest to follow-up of the experimental subjects for whom complete data were available on the nine pretested widowhood functioning and personality variables still did not yield any significant differences among the three treatments, but the multivariate test for non-zero change from pretest to follow-up was highly significant, $F(10, 32) = 3.79$, $p = .002$. Univariate tests indicated that experimental subjects maintained the significant increase in self-esteem, $F(1, 41) = 18.98$, $p < .001$, and in intensity of grief, $F(1, 41) = 23.7$, $p < .001$, and also maintained their more negative

Table 3
Mean Pre-Post Change in Selected Variables

Variable	Consciousness raising	Confidant	Self-help	Control
Self-esteem	2.53	2.78	1.29	3.82
Intensity of grief	6.46	5.57	1.71	4.58
Health prediction	.20	.14	-.06	-.24
Attitude toward remarriage*	-.47	-.86	-.47	-.65
Other-orientation	-1.83	-1.53	-2.38	.29
Self-orientation	.67	.47	1.31	1.29
Radical versus conservative attitudes toward women	.28	2.13	-.69	1.00

* Change in attitude toward remarriage, by therapist, was $-.04$ for Therapist 1 and -1.18 for Therapist 2.

Table 4
Mean Evaluation Scores by Treatment at
Posttest and Follow-up

Variable	Con- sciousness raising	Con- fident	Self- help
Helpfulness			
Posttest	5.75	5.34	4.39
Follow-up	5.55	5.00	4.06
Educational value			
Posttest	5.75	5.54	4.34
Follow-up	5.85	5.57	4.70

attitudes toward remarriage, $F(1, 41) = 6.94$, $p < .05$. A trend toward increased social role involvement by follow-up was also observed ($p = .08$). Mean changes by treatment are reported in Table 5.

Although differences between treatments did not occur in the change in attitudes toward women from pretest to follow-up among the experimental subjects for whom complete data were available, here again an almost significant change across groups emerged in the multivariate analysis of variance ($p < .07$; see Table 5 for the mean changes in these variables by treatment). The univariate test of change in other-orientation was significant, $F(1, 45) = 5.57$, $p < .05$; at follow-up, subjects held less other-oriented attitudes.

The 3×2 multivariate test of the two group evaluation variables at follow-up did not yield a significant main effect for treatment, although the lowest ratings were still given by the self-help participants and the highest by consciousness raising participants.

Table 5
Mean Change from Pretest to Follow-up in Selected Variables by Treatment

Variable	Consciousness raising	Confident	Self-help
Self-esteem	4.71	3.18	4.56
Intensity of grief	6.65	9.46	6.13
Health prediction	.11	.09	-.06
Attitude toward remarriage	-.88	-.55	-1.13
Social role engagement	1.24	1.64	2.44
Other-orientation	-.61	-1.21	-1.81
Self-orientation	1.00	1.29	-.50
Radical versus conservative attitudes toward women	.28	2.00	-.06

The treatment effect was significant in the univariate test of extent of help from the group, $F(2, 45) = 3.38$, $p < .05$. Group means are reported in Table 4. Examination of the group evaluation means and standard deviations suggested that the reduced level of significance of treatment at follow-up was attributable to the increased variance in rating in almost all cells, rather than to a convergence of treatment means. As in the posttest analysis of these data, the therapist and therapist-treatment interaction effects were not significant.

A 3×2 multivariate analysis of variance was performed on the four behavioral measures obtained at follow-up on the experimental subjects for whom complete data were available. The main effect for treatment was significant, $F(8, 84) = 3.14$, $p < .005$, and were the main effect for therapist, $F(4, 42) = 3.47$, $p < .05$, and the therapist-treatment interaction, $F(8, 84) = 4.45$, $p < .001$. Group means for these variables appear in Table 4.

Univariate tests yielded a significant treatment effect for both quality of life change, $F(2, 45) = 3.36$, $p < .005$, and number of group members contacted, $F(2, 45) = 7.03$, $p < .005$. Pearson's correlation of life change scores by the two raters was .89 ($p < .001$). The most positive life changes occurred in the consciousness raising groups, and the least positive changes, in the self-help groups. The content analysis of the types of reported changes across all subjects appears in Table 7. Changes occurred in all spheres of living (except religion) but predominated in the area of mental health. The most frequently

Table 6
Group Means for Behavioral Measures at Follow-up

Variable	Therapist	Consciousness raising	Confidant	Self-help
No. persons contacted	1	1.82	4.12	.43
	2	1.33	.50	.80
No. contacts	1	2.82	13.00	22.57
	2	5.33	1.00	1.50
Group meetings	1	.82	1.25	.14
	2	.00	.00	.90
Life change	1	4.09	2.88	3.14
	2	4.44	4.17	3.40

reported changes were reduced feelings of unique experience; increased self-confidence; more positive future outlook; the incorporation of help from the group; increased social contacts; return to school; and on the negative side, contraction of personal illness. On the average, confidant subjects contacted the most group members; self-help subjects contacted the fewest. At follow-up, four of the six treatment groups planned to continue meeting on their own.^a

Univariate tests on the behavioral measures substantiated the multivariate therapist effect. Both number of members contacted, $F(1, 45) = 6.97$, $p < .01$, and total number of contacts, $F(1, 45) = 4.87$, $p < .05$, varied with the therapist. Subjects in more of the first therapist's groups had higher rates of contact during follow-up. (Because of the differential starting dates, these groups had a 2-week longer follow-up.) Univariate tests for the therapist-treatment interaction were significant for the number of members contacted, $F(2, 45) = 8.04$, $p < .001$, and the number of spontaneous group meetings, $F(2, 45) = 5.4$, $p < .01$, and a trend occurred for total number of contacts ($p = .06$). Examination of the group means revealed that discrepant outcomes occurred in the two confidant groups. One therapist's group members experienced the highest levels of follow-up contact in this treatment; the other's experienced the lowest levels here.

Discussion

The comparison of groups at the pretest indicated that an approximation of random

assignment of subjects was achieved. The treatment groups were somewhat successful in facilitating greater change than that experienced in the control group. Experimental subjects were more likely to give up the view that others' needs are more important than their own. That this change was not limited to consciousness raising participants suggests that a group experience provided a new reference group of widowed women that facilitated new norms that were competing effectively with married group norms in the population. This may be one advantage of group as opposed to individual interventions with the widowed.

The more positive health predictions among experimental subjects is particularly noteworthy in light of the known physical impairment of the widowed (and the actual illness of several subjects), the fact that no medical intervention was offered, and the fact that this significant difference was based on change in a 4-point scale. The index of actual physical complaints did not change, perhaps because the ailments addressed in this variable all commonly occur early in bereavement, whereas many subjects had been widowed for relatively long periods of time. Or perhaps the change in prediction of future health is better understood as an indication of morale. Dunkle (Note 6) found that changing health was related to a change in morale among elderly women but not among elderly men. Parkes (1970) claimed that widows'

^a Subsequent correspondence documented that these plans were carried out.

Table 7
*Life Changes Reported at Follow-up by
 Experimental Subjects*

Area	n
Physical and mental health	
Group reduced feeling that widow's situation was unique	9
More self-confidence	9
More positive future outlook	7
Received help from the group	6
Personal illness	5
Greater decision-making ability	3
Reduced grief	3
Increased goal directedness	2
More insight into self	2
Experienced feelings of inadequacy in relationships with men	2
Group increased feeling of belonging	1
Greater appreciation of life	1
Reduced loneliness	1
Home and family	
Improved family relationships	3
Illness in family	3
Moved	2
Death in family	2
Reunited with family members	2
Separated from family members	1
Renewed interest in improving home environment	1
Less concern with family problems	1
Increased pride in family member	1
Social activities	
More social contacts (nonspecific)	6
Enjoyed group meetings	4
Travel	4
Positive experiences with men	4
Joined a club	3
Doing more things (nonspecific)	2
Fewer social contacts	1
A friend died	1
Assumed jury duty	1
New involvement in charitable work	1
Received community recognition award	1
Employment and education	
Went back to school	5
Got a new job	3
Quit or fired from job	3
Experienced job frustration	2
Received a scholarship	1
Quit school	1

Note. n = 5 for no change. N = 51.

global self-reports of *current* poor health were related more to feelings of anger and irritability than to physical symptoms per se, although data were not presented to document this finding.

The treatment groups also were successful as catalysts for social interaction from pretest to follow-up. High rates of contact among group members were generated. Treatment plans for continued meetings even after follow-up attest to the program's effectiveness in providing an antidote to loneliness.⁷ Psychotherapy outcome studies seldom include extensive follow-up, and when they do, results are often discouraging. Yet in this study, treatment gains were maintained for several months.

The simultaneous gains in self-esteem and intensity of grief merit discussion. Widowed persons frequently complain that others are unwilling to listen to their negative feelings. Unfinished "grief work" (Caplan, 1964) may be common in this population, despite years of widowhood. Supported by a group of peers, there may be nothing contradictory about the simultaneous expression of grief and increased positive regard for oneself. It is possible that the increased intensity reported reflects a new acceptance of negative emotional reactions to widowhood and a lessening of their denial. In any event, the apparent increase in grief amidst an array of other positive changes casts doubt on the rationale of treatment modalities for widowed persons that focus primarily on the *reduction* of feelings of sorrow, anger, depression, guilt, and so on.

Of the three types of widows' groups pioneered in this research, the most consistently effective method was the widows' consciousness raising group, and treatment outcomes were remarkably similar across therapists. The high degree of structure in the consciousness raising groups, which guaranteed opportunity

⁷ In this regard, it is interesting that a particular task of the confidant groups presumed to require a fairly high level of intimacy—the sharing of telephone numbers in order to plan a mutual activity outside the group—became redundant, as telephone numbers were spontaneously exchanged by almost all groups by midtreatment.

nities for all members to participate actively, may have contributed to their success. Alternatively, the content of these groups may have been particularly therapeutic. The focus on women and sex role oppression provided a natural external target for anger. This emotion has long been observed to be facilitative in the treatment of depression.

The self-help format on the whole was least effective, although no between-treatment differences emerged on a number of variables reflecting positive change. Most self-help groups do not have a professional leader. It is possible that the presence of a therapist in these groups was counterproductive, despite, or perhaps because of, her limited role.

Response to the confidant groups was more variable. Both therapists found these groups the most difficult to lead, as departures in the session plans were often necessitated by absences or by a discussion of the members' objections. The very idea of becoming particularly well acquainted with only one widow appeared to foster apprehension. (The average confidant subject in fact contacted *more* members during follow-up than subjects in either of the other treatments.) The confidant group strategy might be more effective if limited to widows who have *no* confidant prior to treatment.

A major finding of this research was that substantial change occurred in *all* groups including the control group. The promise of an extended small group experience in 2 months time may itself have had a therapeutic effect on the waiting list controls, especially in view of the reduced opportunities for social interaction and the scarcity of helping resources in widowhood. This interpretation is consistent with my clinical observation that an enthusiastic, eager quality in the initial meeting of the waiting list discussion group replaced the more depressive tone in the initial meetings of the earlier groups. It may be worthwhile to explore the potential therapeutic benefits of waiting periods with other clinical populations.

The gains of the control group may also reflect a positive change bias in those who tolerated the waiting period and returned. However, this is unlikely as a total explanation,

since the percentage of control subjects who did not return (35.8%) is not that different from the percentage of subjects assigned to a treatment group who never attended (29%). It could also be argued that waiting list subjects who improve on their own should be *less* likely to return for treatment.

Attention should be addressed to the subject characteristics that increase the probability of enrollment in a therapeutic program for widows and those that are associated with the greatest therapeutic impact. The fact that widows of all ages and duration of widowhood enrolled in the program refutes the views that the stresses of widowhood are limited to a particular age group and that only the recently widowed need help. Clearly, the strategy of placing widows varying widely in age and duration of widowhood⁸ in the same group was not a hindrance to the program's effectiveness, and it may have been facilitative. Experienced widows probably functioned as role models for the newcomers, and the diversity of the members probably served to remind the widow of available options in her own life-style decisions.

Compared to widows nationally, subjects in this research overrepresented the Jewish faith, high educational achievement, and prior upper-income status. The location of the program on a university campus and the required fee may have contributed to the observed sample bias. The generalizability of the results awaits replication with other widowed samples.

A number of dependent variables in this research did not evolve as effective descriptors of the personal changes that accompany participation in a social service for widows (locus of control,⁹ life satisfaction, frequency of physical complaints, attitude toward widowhood, and the subset of items from the radical versus conservative Attitudes Toward Women scale). Some of the measures may have been too global to reflect the changes

⁸ Few extremely recent widows participated.

⁹ The I-E scale may have limited applicability to an elderly population. Subject complaints about the irrelevance of items were common.

that occurred. Subsequent research might focus on the discrete types of life changes reported by subjects on the open-ended question, that is, the feeling of unique experience, attitude toward the future, self-confidence, extent of social contacts, and so on.

This research has addressed the question of how best to facilitate therapeutic change among widows using a group format. It provides the best documentation to date that psychological and social change can be fostered by innovative programs for the widowed. Mental health professionals as well as widowed persons should be able to assume an attitude of hopefulness instead of resignation with respect to the circumstances of widowhood. Some subjects experienced far-reaching changes in several areas of their lives during the 6 months from the time they read about the program to the follow-up after their small group experience. A question for the future is how to predict which persons will profit most from a widows' group.

Reference Notes

1. Holmes, T. H., & Masuda, M. *Life change and illness susceptibility*. Paper presented at the meeting of the American Association for the Advancement of Science, Chicago, December 1970.
2. Antoniuk, H. Personal interview, San Diego, February 1973.
3. Van Coevering, V. *Exploring group counseling as a technique for ameliorating morbidity, mortality, and lowered life satisfaction of widowhood*. Paper presented at the meeting of the Gerontological Society, Portland, Oregon, November 1974.
4. Parkes, C. M. Lecture addressed to the Los Angeles County—University of Southern California Medical Center Department of Psychiatry, Los Angeles, October 1973.
5. Freeman, V. Personal interview, March 1973.
6. Dunkle, R. E. *Role discontinuity as it effects changes in adjustment of women to aging*. Paper presented at the meeting of the Gerontological Society, Miami Beach, November 1973.
7. Lopata, H. Z. Personal communication, October 1975.

References

- Abrahams, R. B. Mutual help for the widowed. *Social Work*, 1972, 17(5), 54-61.
- Adams, B. N. The middle-class adult and his widowed or still-married mother. *Social Problems*, 1968, 16, 50-59.
- Barrett, C. J. Women in widowhood: Review essay. *Signs: Journal of Women in Culture and Society*, 1977, 2, 856-868.
- Bellin, S. S., & Hardt, R. H. Marital status and mental disorders among the aged. *American Sociological Review*, 1958, 28, 155-162.
- Berardo, F. M. Survivorship and social isolation: The case of the aged widower. *Family Coordinator*, 1970, 19, 11-25.
- Bock, E. W., & Webber, I. L. Suicide among the elderly: Isolating widowhood and mitigating alternatives. *Journal of Marriage and the Family*, 1972, 34, 24-31.
- California Department of Public Health, California Health Survey. *Health in California*. Sacramento: State Printing Office, 1958.
- Caplan, G. *Principles of preventive psychiatry*. New York: Basic Books, 1964.
- Carter, H., & Glick, P. C. *Marriage and divorce: A social and economic study*. Cambridge, Mass.: Harvard University Press, 1970.
- Clayton, P. Mortality and morbidity in the first year of widowhood. *Archives of General Psychiatry*, 1974, 30, 747-750.
- Cosneck, B. J. A study of the social problems of aged widowed persons in an urban area. (Doctoral dissertation, Florida University, 1966) *Dissertation Abstracts International*, 1966, 27, 3525A. (University Microfilms No. 67-3469.)
- Gorer, G. *Death, grief and mourning*. New York: Doubleday, 1965.
- Gump, J. P. Sex-role attitudes and psychological well-being. *Journal of Social Issues*, 1972, 28(2), 79-92.
- Heyman, D. K., & Gianturco, D. T. Long-term adaptation by the elderly to bereavement. *Journal of Gerontology*, 1973, 28, 359-362.
- Hurvitz, N. Peer self-help psychotherapy groups and their implications for psychotherapy. *Psychotherapy: Theory, Research and Practice*, 1970, 7, 41-49.
- Kraus, A. A., & Lilienfeld, A. M. Some epidemiologic aspects of the high mortality rate in the young widowed group. *Journal of Chronic Disease*, 1959, 10, 207-217.
- Lewis, A. A., & Berns, B. *Three out of four wives*. New York: MacMillan, 1975.
- Lopata, H. Z. *Widowhood in an American city*. Cambridge, Mass.: Schenkman, 1973.
- Lowenthal, M. F. Social isolation and mental illness in old age. *American Sociological Review*, 1964, 29, 54-70.
- Lowenthal, M. F., & Haven, C. Interaction and adaptation: Intimacy as a critical variable. *American Sociological Review*, 1968, 33, 20-30.
- Maddison, D., & Viola, A. The health of widows in the year following bereavement. *Journal of Psychosomatic Research*, 1968, 12, 297-306.
- Marris, P. *Widows and their families*. London: Routledge & Paul, 1958.
- Miles, H. S., & Hays, D. R. Widowhood. *American Journal of Nursing*, 1975, 75(2), 280-282.
- Neugarten, B. L., Havighurst, R. J., & Tobin, S. S.

- The measure of life satisfaction. *Journal of Gerontology*, 1961, 16, 134-143.
- Nuckols, R. C. Widows study. *JSAS Catalog of Selected Documents in Psychology*, 1973, 3, 9. (Ms. No. 296, 297, 298)
- Palmore, E. B., Stanley, G. L., & Cormier, R. H. *Widows with children under social security*. (Research Rep. 16) Washington, D.C.: U.S. Department of Health, Education and Welfare, Social Security Administration, 1963.
- Parkes, C. M. Effects of bereavement on physical and mental health: A study of the medical records of widows. *British Medical Journal*, 1964, 2, 274-279.
- Parkes, C. M. The first year of bereavement. *Psychiatry*, 1970, 33, 444-467.
- Parkes, C. M., Benjamin, B., & Fitzgerald, R. G. Broken heart: A statistical study of increased mortality among widowers. *British Medical Journal*, 1969, 1, 740-743.
- Parkes, C. M., & Brown, R. J. Health after bereavement. *Psychosomatic Medicine*, 1972, 34, 449-461.
- Robertson, N. C. The relationship between marital status and the risk of psychiatric referral. *British Journal of Psychiatry*, 1974, 124, 191-202.
- Rosenberg, M. The association between self-esteem and anxiety. *Journal of Psychiatric Research*, 1962, 1(2), 135-152.
- Rosenfeld, L. S., Katz, J., & Donabedian, S. *Medical care needs and services in the Boston Metropolitan area*. Boston: United Community Services of Metropolitan Boston, 1957.
- Rosenfeld, L. S., Mott, F. D., & Taylor, M. D. Health services for the aging in Saskatchewan. In, *Illness and health services in an aging population*. St. Louis, Mo.: Second International Gerontological Congress, 1951.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 1966, 80(1, Whole No. 609).
- Schlesinger, B. The widowed as a one-parent family unit. *Social Science*, 1971, 46, 26-32.
- Segal, B. E. Suicide and middle age. *Sociological Symposium*, 1969, 3, 131-140.
- Silverman, P. R. Services for the widowed during the period of bereavement. In, *Social Work Practice: Proceedings*. New York: Columbia University Press, 1966.
- Silverman, P. R. The widow as a caregiver in a program of preventive intervention with other widows. *Mental Hygiene*, 1970, 54, 540-547.
- Simon, A. Physical and sociopsychologic stress in the geriatric mentally ill. *Comprehensive Psychiatry*, 1970, 11, 242-247.
- Spence, J. T., & Helmreich, R. The Attitudes Toward Women Scale: An objective instrument to measure attitudes toward the rights and roles of women in contemporary society. *JSAS Catalog of Selected Documents in Psychology*, 1972, 2, 66-67. (Ms. No. 153)
- Stein, Z., & Susser, M. Widowhood and mental illness. *British Journal of Preventive and Social Medicine*, 1969, 23, 106-110.
- United States Bureau of the Census, *Census of the Population: 1970. Characteristics of the Population*, 1970, 1, 42.
- Woods, B. Widows: More than loneliness. (Interview with D. Horowitz) *Los Angeles Times*, August 20, 1973, Pt. 4, 1.
- Woolsey, T. D. *Estimates of disabling illness prevalence in the U.S., based on the current population survey of February 1949 and September 1950*. (U.S. Public Health Service Rep. 181) Washington, D.C.: Government Printing Office, 1952.

Received August 11, 1976 ■

Reduction of Test Anxiety Through Cognitive Restructuring

Marvin R. Goldfried

State University of New York at Stony Brook

Marsha M. Linehan and Jean L. Smith

The Catholic University of America

This collaborative clinical outcome study compared two procedures for reducing test anxiety with a waiting list control. In the first, systematic rational restructuring, participants were trained to realistically reevaluate imaginatively presented test-taking situations. In the second, a prolonged exposure condition, the same hierarchy items were presented but with no instructions for coping cognitively. On the basis of questionnaire measures of test anxiety, greater anxiety reduction was found in the systematic rational restructuring condition, followed by the prolonged exposure group, with no changes for the waiting list control. Only those in the rational restructuring condition reported a significant decrease in subjective anxiety when placed in an analogue test-taking situation. Participants in the restructuring condition also reported greater generalized anxiety reduction in social-evaluative situations. Within the broader context of cognitive behavior therapy, the results of the present investigation add to the increasing number of outcome studies indicating that the cognitive reappraisal of anxiety-provoking situations can offer a markedly effective treatment procedure for the reduction of anxiety.

The incorporation of cognitive variables within behavior therapy represents a clear and unmistakable trend. Much of the current work in this area has been based on the clinical observation of Ellis (1962), who has argued that modification of inappropriate expectations and beliefs could lead to behavior change. Until recently, however, Ellis' rational-emotive therapy had relatively little impact on the behavioral movement. The difficulty of fitting Ellis' approach into a behavioral orientation has been due, in part, to

the lack of clear therapeutic guidelines as well as to the absence of an empirical data base for its clinical effectiveness. This situation is clearly changing, and steps are currently being taken to incorporate many of Ellis' concepts and procedures into the field of cognitive behavior therapy (Beck, 1976; Goldfried & Davison, 1976; Goldfried, Decenteeo, & Weinberg, 1974; Mahoney, 1974; Meichenbaum, 1977). A number of outcome studies have appeared in the literature, demonstrating that speech anxiety (Meichenbaum, Gilmore, & Fedoravichous, 1971; Trexler & Karst, 1972), interpersonal anxiety (DiLoreto, 1971; Kanter & Goldfried, Note 1), unassertive behavior (Thorpe, 1975; Wolfe & Fodor, in press; Linehan, Goldfried, & Goldfried, Note 2), and test anxiety (Holroyd, 1976; Meichenbaum, 1972; Osarchuk, 1976) can be reduced by intervention procedures that focus on training individuals to modify their unrealistic belief systems.

Like much of the early behavior therapy research in general, therapeutic attempts at the reduction of test anxiety have focused

The research described in this article was supported in part by Grant MH24327 from the National Institute of Mental Health.

The authors are grateful to Ed Harpin for his assistance in carrying out the study, to Tom Holman for his suggestions regarding the statistical analyses, and to Gerald C. Davison and Sharon Foster for their comments on an earlier version of this article. The second author is now at the University of Washington, Seattle, Washington.

Requests for reprints should be sent to Marvin R. Goldfried, Department of Psychology, State University of New York at Stony Brook, Stony Brook, New York 11794.

primarily on the use of systematic desensitization (Wine, 1971). The relevance of a more cognitively oriented approach in the treatment of test anxiety is noted by Wine (1971), whose review suggests that anxious individuals not only experience emotional arousal but also engage in excessive worry about the adequacy of their performance. Based on the assumption that test anxiety may be comprised of both "emotionality" and "worry" components (Liebert & Morris, 1967; Morris & Liebert, 1969), Meichenbaum (1972) developed a treatment package involving cognitive restructuring and modified systematic desensitization. Test-anxious subjects were provided with relaxation training, engaged in discussions of potentially unrealistic beliefs associated with test taking, and were then given practice in coping with imagined test-related situations by means of relaxation and self-instructions to focus on only the test itself. Compared with traditional systematic desensitization, this cognitive modification package produced greater reductions in test anxiety. Although Meichenbaum's study demonstrated that a treatment approach that includes the altering of cognitions can have an effect on reducing test anxiety, a more precise interpretation of these findings is limited by the inclusion of relaxation in the cognitive treatment package. Further, the exposure times for hierarchy presentation were longer for the cognitive modification subjects, since they were instructed to maintain the image while attempting to cope with their anxiety (Meichenbaum, Note 3).

The primary purpose of the present study was to determine whether a treatment procedure using only cognitive restructuring, incorporating the basic tenets of Ellis' therapeutic approach, could be successful in the reduction of test anxiety. The therapeutic procedure used was originally outlined by Goldfried et al. (1974), whose guidelines for the implementation of *systematic rational restructuring* are presented within a social learning framework. The treatment procedure, which is described in greater detail elsewhere (Goldfried & Davison, 1976), essentially involves the use of imaginatively presented hierarchy items to provide individuals with prac-

tice in ferreting out unrealistic concerns and worries, affording them the opportunity to place each situation into a more realistic perspective, and then using their newly acquired skills to reduce anxiety in real-life situations. In the present study, this treatment procedure was compared with an exposure-alone condition, in which the hierarchy items were presented without any instructions or directions for coping cognitively. The prolonged exposure group was included to control for possible extinction effects.

A second purpose of this study was to demonstrate the feasibility of conducting collaborative clinical outcome research with investigators at more than one setting. In an overview of the current status and future direction of psychotherapy research, Bergin and Strupp (1970) discussed the need for such coordinated or collaborative research efforts. As noted by them, a major difficulty associated with collaborative research in psychotherapy has been a general lack of standardization. By focusing on a specific target problem—test anxiety—and using comparable therapeutic techniques and assessment procedures, we wished to demonstrate that such standardization problems could be overcome.

Method

Participants

Participants were 15 men and 21 women who responded to advertisements for treatment of test anxiety at the State University of New York at Stony Brook and The Catholic University of America. They ranged in age from 18 to 49 years and were not being seen in therapy elsewhere. Of the 42 subjects who originally began the study, 6 were eliminated for failure to complete the treatment. The attrition was distributed among the three conditions as follows: 2 in rational restructuring, 1 in prolonged exposure, and 3 in waiting list control.

Procedure

After responding to local advertisements, interested individuals were sent a questionnaire battery, a program description, and a consent form. Persons who returned the consent form and questionnaire battery and who were available at scheduled therapy times were assigned by a within-sample matching technique (see Goldstein, Heller, & Sechrest,

1966) to systematic rational restructuring ($n=12$), prolonged exposure ($n=13$), or waiting list ($n=11$) conditions. Participants were then seen in person and were given a pretreatment analogue examination and an associated assessment battery. Therapy, administered in groups over six 1-hour sessions, was standardized across the two treatment conditions. Participants were reassessed within the week following the termination of therapy and at a 6-week follow-up.

The study was carried out in two waves. Approximately half of the individuals were seen at Stony Brook, and half participated 1 year later at Catholic University. Detailed therapy manuals, transcripts of therapy sessions, audiotapes of pilot sessions, and phone calls were exchanged to ensure maximum comparability between the two locations. All other aspects of the study were identical.

Measures

Two classes of measures were used in evaluating the results of the treatment procedures: a questionnaire battery and subjective reports of anxiety preceding an analogue examination situation. The questionnaire battery consisted of several scales designed to measure test anxiety: the S-R Inventory of Anxiousness (Endler, Hunt, & Rosenstein, 1962) for situations "taking an important exam" and "taking a weekly quiz"; the Suinn Test Anxiety Behavior Scale (Suinn, 1969); the Achievement Anxiety Test derived by Alpert and Haber (1960), which assesses facilitating as well as debilitating anxiety in examination situations; and the Test Anxiety Questionnaire (Mandler & Sarason, 1952), using Liebert and Morris' (1967) Worry and Emotionality subscores. Measures included to assess generalization of therapy effects to areas other than test anxiety were the S-R Inventory of Anxiousness for the situations "giving a speech," "going to a party," and "interviewing for a job"; the Fear of Negative Evaluation and Social Avoidance and Distress scales (Watson & Friend, 1969); and the Trait scale of the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970). Participants also responded to a question measuring their expectations for benefit from therapy, using a 5-point scale ranging from "0% chance of success" to "100% chance of success."

After completing the questionnaire, participants were asked to take part in an actual "examination" situation. The examination was administered in small groups and consisted of the Wonderlic Personnel Test (Wonderlic, 1959) and the Digits Forward, Digits Backward, and Digit Symbol tests taken from the Weschler Adult Intelligence Scale. Participants were told that the test they were taking was selected from standardized IQ tests and that the results would be posted beside their names. Immediately following these instructions and before taking the exam, participants were administered the State scale of the State-Trait Anxiety Inventory (Spiel-

berger et al., 1970); the Anxiety Differential (Hush & Alexander, 1963); and the Experiences Questionnaire, which consists of Taylor Manifest Anxiety Scale items analyzed into Worry and Emotionality subscores (Morris & Liebert, 1969). Posttesting included the identical questionnaire battery and pre-exam measures, and the examination situation was composed of alternate forms of the tests administered during the pretest. At a 6-week follow-up, participants in the two treatment conditions were sent the same questionnaire battery plus forms for evaluating the program and their respective therapists. Because of ethical considerations, individuals in the waiting list control were offered therapy immediately following the posttest and were therefore not included in the follow-up assessment.

Therapy

Therapy sessions in the systematic rational restructuring and prolonged exposure conditions were conducted in groups that met weekly for six 1-hour sessions. Participants who missed sessions listened to audiotapes of the missed session, carried out the procedure described on the tape, completed in-session practice sheets, and returned these with homework assignments to an assistant. As a means of assessing the credibility of each treatment rationale, participants were asked to estimate the amount of success that they expected from therapy at the conclusion of the first session.

Systematic rational restructuring condition. The therapy program followed the systematic rational restructuring procedure outlined in Goldfried and Davison (1976). During the first session, the treatment rationale and therapy procedures were outlined, and participants received practice in imagining situations. During Sessions 2-6, the participants were presented with a standard 15-item hierarchy constructed on the basis of Stony Brook participants' pretest responses to the Suinn Test Anxiety Behavior Scale (Suinn, 1969). Three situations were presented in each session, and participants were instructed to imagine themselves in the situation throughout the presentation of each item and to attempt to reduce their anxiety by means of rational restructuring. Each item was presented for a total of four 1-minute trials. Immediately following each trial participants were instructed to record their self-defeating thoughts (e.g., "I'm going to fail this test, and then everyone's going to think I'm stupid"), their rational reevaluation (e.g., "Chances are I probably won't fail. And even if I do, people probably won't think I'm stupid. And even if they do, that doesn't mean that I am stupid"), and their anxiety levels before and after reevaluating. A brief group discussion followed the fourth presentation of each item. Participants were instructed to practice their reevaluation skills in vivo and were provided with homework sheets that served as the basis for discussion at the outset of the following session.

Prolonged exposure condition. A social-learning-based rationale was presented to subjects in this condition, emphasizing the importance of habituation and extinction in the reduction of anxiety. The first session of the treatment was identical in general format to that in the systematic rational restructuring condition. The same hierarchy presentation used for the restructuring condition was used; during scene presentations in this condition, however, participants were instructed to focus on their emotional reactions. In-session record forms were provided for the recording of feelings noted at the beginning and end of each imaginal presentation, and a group discussion was held after all four presentations of each item. Homework assignments, discussed at the beginning of each session, consisted of attending to their anxiety reactions in everyday situations.

Waiting list control. Participants in this condition were informed that because of the limited therapeutic time available, there would be a brief delay before they could be seen in treatment. Participants received the same pretest and posttest assessment battery as was administered to individuals in the two therapeutic contact conditions, and they were then offered therapy following the posttesting.

Therapists

The first and second authors served as therapists at their respective universities. A detailed therapy manual was used, hierarchy scenes were written out fully and were delivered verbatim, and transcripts and tapes of pilot sessions were exchanged to minimize any differences between the therapists. Both in-person meetings and frequent phone consultations during the progress of the experimental sessions insured that therapy and subject problems were handled in the same way by both therapists, and that the assessment procedures were administered in a comparable fashion. Each therapist had contact with approximately the same number of participants in each treatment condition.

Results

Separate univariate analyses of variance conducted on each dependent variable indicated no significant pretest differences among the three groups on any of the measures used in the study. Inasmuch as comparisons between therapists failed to reveal any main effect differences, the data from the two universities were combined in all subsequent analyses.¹

Pre-Post Treatment Effects

Questionnaire battery. The results of one-way analyses of covariance on each of the

questionnaire variables are presented in Table 1. On variables designed to measure test anxiety, there were significant treatment effects on the S-R Inventory of Anxiousness for both the "exam" and "quiz" situations, the Suinn Test Anxiety Behavior Scale, and the Achievement Anxiety Test Debilitating Anxiety scale. Newman-Keuls tests for differences between adjusted means revealed that participants in the systematic rational restructuring condition reported significantly less anxiety in exam situations, less anxiety on the Suinn scale, and less debilitating anxiety than did participants in the prolonged exposure condition. Both groups reported less anxiety on these three measures than did individuals in the waiting list control. Participants in the two therapy conditions did not differ from each other on S-R inventory reports of anxiety in quiz situations, but both reported significantly less quiz anxiety than did those on the waiting list. There were no significant differences among groups on the Facilitating Anxiety scale or on the Worry or Emotionality subscales of the Test Anxiety Questionnaire.

Significant generalization effects were found on the S-R Inventory of Anxiousness for "party" and "job interview" situations and on the Fear of Negative Evaluation and the Social Avoidance and Distress scales. Newman-Keuls comparisons of adjusted means indicated that participants in systematic rational restructuring reported less Fear of Negative Evaluation and Social Avoidance and Distress than those in either the prolonged exposure or the waiting list control conditions, which did not differ significantly from each other. On the S-R Inventory of Anxiousness for both party and job interview situations, participants in the systematic rational restructuring condition reported less anxiety than those on the waiting list; no difference was found between the restructuring and prolonged exposure groups. There were no significant differences among groups on the Trait scale of the State-Trait Anxiety

¹ The means and standard deviations for pretest, posttest, and follow-up assessments are available from the first author.

Table 1

One-Way Analyses of Covariance and Adjusted Means at Posttesting

Measure	<i>F</i>	Adjusted treatment <i>M</i>		
		Systematic rational restructuring	Prolonged exposure	Waiting list
Questionnaire Battery: Test Anxiety				
S-R inventory: exam	10.10***	35.67 _a	41.38 _b	46.55 _c
S-R inventory: quiz	6.71**	30.57 _a	34.67 _a	42.40 _b
Suinn Test Anxiety Behavior Scale	15.81***	116.55 _a	149.06 _b	171.86 _c
Achievement Anxiety Test: debilitating	14.11***	27.23 _a	31.27 _b	37.52 _c
Achievement Anxiety Test: facilitating	1.70	21.11 _a	21.79 _a	18.85 _a
Test Anxiety Questionnaire: worry	2.75	37.52 _a	41.78 _a	44.96 _a
Test Anxiety Questionnaire: emotionality	2.22	34.01 _a	38.18 _a	38.14 _a
Questionnaire Battery: Generalization				
S-R inventory: speech	1.03	33.93 _a	35.78 _a	38.34 _a
S-R inventory: party	3.49*	25.91 _a	31.40 _{a,b}	33.62 _b
S-R inventory: job	3.64*	31.05 _a	36.79 _{a,b}	39.46 _b
Fear of Negative Evaluation	6.37**	10.98 _a	15.79 _b	16.55 _b
Social Avoidance & Distress	8.78***	5.04 _a	9.66 _b	13.09 _b
Trait Anxiety scale of the STAI	1.87	41.31 _a	46.52 _a	47.15 _a
Preexamination anxiety				
State Anxiety scale of the STAI	2.72	35.46 _a	39.98 _a	42.43 _a
Anxiety Differential	<1	62.23 _a	62.41 _a	63.36 _a
Experiences Questionnaire: worry	4.62*	7.52 _a	9.08 _{a,b}	11.33 _b
Experiences Questionnaire: emotionality	3.08	6.91 _a	8.72 _a	11.79 _a

Note. Means in the same row with different subscripts differ from each other at least at the .05 level. For analyses of covariance, $df = 2, 32$. STAI = State-Trait Anxiety Inventory.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Inventory or the "speech" situation on the S-R inventory.

Focusing specifically on the direction of change, analyses of pre-post within-group differences by t tests (see Table 2) indicated a significant improvement in systematic rational restructuring across all but one measure designed to measure test anxiety (the Achievement Anxiety Test Facilitating scale) and on all but one measure of treatment generalization (the Trait scale of the State-Trait Anxiety Inventory). In contrast, the prolonged exposure participants improved significantly on only three variables; the Achievement Anxiety Debilitating scale, the Worry subscale of the Text Anxiety Questionnaire, and the Social Avoidance and Distress Scale. No pre-post differences were found on any measure for the waiting list control. Although

these within-group changes cannot be interpreted in terms of difference *between* the two treatment groups, they are nonetheless consistent with the overall superiority of the systematic rational restructuring group.

Examination situation. One-way analyses of covariance were carried out separately on posttest scores for the State scale of the State-Trait Anxiety Inventory, the Anxiety Differential, and the Worry and Emotionality subscales of the Experiences Questionnaire. As can be seen in Table 1, a significant main effect for treatment was found only on the Worry subscale. Newman-Keuls individual comparisons indicated that participants who received rational restructuring reported less preexam worry than those in the waiting list condition. No significant difference was found

between the systematic rational restructuring and the prolonged exposure conditions.

Analyses of pre-post differences, summarized in Table 2, indicated a significant reduction in subjective anxiety for the systematic rational restructuring condition on all but one variable, the Anxiety Differential. There was no significant pre-post improvement on any of the four measures for either the prolonged exposure or waiting list control.

Follow-up Evaluation

Analyses of covariance were carried out to test whether treatment differences were maintained on the questionnaire battery measures at follow-up 6 weeks after the termination of therapy. Results indicated that systematic rational restructuring participants reported significantly less anxiety than prolonged ex-

posure participants on the Suinn Test Anxiety Behavior Scale, $F(1, 22) = 8.25, p < .01$; the S-R Inventory of Anxiousness party situation, $F(1, 22) = 5.09, p < .05$; and the Social Avoidance and Distress Scale, $F(1, 22) = 5.71, p < .05$. In addition, there was a trend at the .10 level indicating less debilitating anxiety on the Achievement Anxiety Test in the systematic rational restructuring group as compared to the prolonged exposure group. There were no other significant differences between the two groups.

At the follow-up assessment, participants were asked to rate the therapy that they had received on several dimensions. Analyses by t tests indicated that participants in the systematic rational restructuring condition, compared to those in the prolonged exposure condition, were generally more satisfied with changes in themselves, $t(22) = 2.68, p < .05$.

Table 2
Within-Group Mean Differences from Pretesting to Posttesting

Measure	Condition		
	Systematic rational restructuring	Prolonged exposure	Waiting list
Questionnaire battery: Test anxiety			
S-R inventory: exam	8.50***	2.46	-2.00
S-R inventory: quiz	8.33**	2.93	-3.18
Suinn Test Anxiety Behavior Scale	52.17***	19.61	1.00
Achievement Anxiety Test: debilitating	7.75***	3.15*	-2.82
Achievement Anxiety Test: facilitating	-1.58	1.39	.09
Test Anxiety Questionnaire: worry	9.33**	5.54**	1.73
Test Anxiety Questionnaire: emotionality	5.25*	1.85	1.10
Questionnaire battery: Generalization			
S-R inventory: speech	5.08**	3.00	.72
S-R inventory: party	8.33**	2.31	-2.09
S-R inventory: job	8.17*	3.08	-.73
Fear of Negative Evaluation	6.16***	1.38	.18
Social Avoidance & Distress	6.59**	4.00*	-1.27
Trait Anxiety scale of the STAI	3.67	2.93	-1.00
Preexamination anxiety			
State Anxiety scale of the STAI	11.58**	4.23	3.63
Anxiety Differential	-1.08	-1.16	-1.64
Experiences Questionnaire: worry	3.83**	1.00	-.27
Experiences Questionnaire: emotionality	3.25*	1.00	.00

Note. Significance of change is based on within-group t tests. STAI = State-Trait Anxiety Inventory.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

The groups did not differ in their ratings of the helpfulness of therapy for test anxiety or on their estimate of the generalization effects of the therapies.

Expectancies and Liking of Therapist

During the pretest, participants were asked to rate their expectations for becoming less anxious in test-taking situations. This initial expectancy rating was used as one of the variables for the within-sample matching procedure. A check on this matching procedure was conducted by a one-way analysis of variance, which failed to reveal any significant differences among the three conditions. To check whether the two therapy rationales were equally credible, participants again rated their expectancy for success on the same 5-point scale after the first session. The absence of a significant *t*-test difference suggests that the two groups were comparable with regard to the nonspecific effects associated with expectancy of improvement and treatment demand characteristics.

At follow-up, participants in the two contact conditions were asked to rate their therapist on the following 5-point bipolar scales: incompetent-competent, unlikable-likable, not understanding-understanding, aloof-warm, and uncomfortable-comfortable. The ratings from the two conditions were all toward the positive pole and were not significantly different from each other on any of the scales. These findings suggest that differences between the two treatments cannot be attributed to nonspecific therapist differences while conducting the two treatments.

Discussion

The results of this study revealed a consistent pattern: Based on questionnaire measures of test anxiety, participants in the systematic rational restructuring condition experienced greater anxiety reduction, followed by those having undergone prolonged exposure to the same hierarchy items, whereas the waiting list control did not change. Only participants in systematic rational restructuring reported a significant pre-post decrease in subjective anxiety before an analogue test-taking situation. In addition to the reduction of test anxiety, participants in the restructur-

ing condition also reported greater generalization of anxiety reduction in social-evaluative situations.

Although the systematic rational restructuring procedure was found to be the most effective of three conditions, exposure alone also produced significant anxiety reduction. These findings corroborate the results of several other studies that have demonstrated that hierarchy exposure alone, particularly if it is prolonged or repeated, can be an effective therapeutic procedure for anxiety reduction (e.g., D'Zurilla, Wilson, & Nelson, 1973; Goldfried & Goldfried, 1977; Malleon, 1959). These therapy-analogue findings are consistent with the animal literature on extinction of avoidance via exposure and response prevention (Wilson & Davison, 1971). As indicated in the present study, however, active attempts to cope with one's anxiety by means of cognitive reappraisal adds to any anxiety reduction associated with extinction or habituation and greatly facilitates generalization to nontreated targets.

This study has also demonstrated that collaborative research with investigators at different settings is possible and that it can be a useful approach to increasing the number of participants available in clinical outcome research and facilitating the generalizability of the findings. Such collaborative research efforts are feasible provided that close coordination and appropriate methodological precautions are taken, such as insuring the standardization of treatment procedures, comparability of participants, and equivalence of outcome measures.

Although Meichenbaum's (1972) cognitive modification procedure for reducing of test anxiety consisted of a treatment package containing rational-emotive therapy, self-instructions, and relaxation components, our results reveal that cognitive restructuring alone is effective. It should be noted, however, that the cognitive restructuring procedure used by Meichenbaum is somewhat different from the intervention procedure used in the present study. Although Meichenbaum included a rational-emotive therapy component, the self-instructions used during hierarchy presentation involved self-statements to encourage participants to pay attention to

the task at hand. By contrast, participants in rational restructuring were taught to tune in to any of their unrealistic concerns and worries and to put them into a more realistic perspective. Thus, the results of the present study add to the increasing number of outcome studies indicating that the cognitive reappraisal of anxiety-provoking situations is a markedly effective treatment procedure for the reduction of anxiety.

Reference Notes

1. Kanter, N. J., & Goldfried, M. R. *Relative effectiveness of rational restructuring and self-control desensitization in the reduction of interpersonal anxiety*. Unpublished manuscript, 1976.
2. Linehan, M. M., Goldfried, M. R., & Goldfried, A. P. *Assertion therapy: Skill training or cognitive restructuring?* Unpublished manuscript, 1977.
3. Meichenbaum, D. Personal communication, August 1976.

References

- Alpert, R., & Haber, R. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, 1960, 61, 207-215.
- Beck, A. T. *Cognitive therapy and the emotional disorders*. New York: International Universities Press, 1976.
- Bergin, A. E., & Strupp, H. H. New directions in psychotherapy research. *Journal of Abnormal Psychology*, 1970, 76, 13-26.
- DiLoreto, A. O. *Comparative psychotherapy: An experimental analysis*. Chicago: Aldine-Atherton, 1971.
- D'Zurilla, T. J., Wilson, G. T., & Nelson, R. A preliminary study of the effectiveness of graduated prolonged exposure in the treatment of irrational fear. *Behavior Therapy*, 1973, 4, 672-685.
- Ellis, A. *Reason and emotion in psychotherapy*. New York: Lyle Stuart, 1962.
- Endler, N. S., Hunt, J. McV., & Rosenstein, A. J. An S-R inventory of anxiousness. *Psychological Monographs*, 1962, 76(17, Whole No. 536).
- Goldfried, M. R., & Davison, G. C. *Clinical behavior therapy*. New York: Holt, Rinehart & Winston, 1976.
- Goldfried, M. R., Decentecio, E. T., & Weinberg, L. Systematic rational restructuring as a self-control technique. *Behavior Therapy*, 1974, 5, 247-254.
- Goldfried, M. R., & Goldfried, A. P. Importance of hierarchy content in the self-control of anxiety. *Journal of Consulting and Clinical Psychology*, 1977, 45, 124-134.
- Goldstein, A. P., Heller, K., & Sechrest, L. B. *Psychotherapy and the psychology of behavior change*. New York: Wiley, 1966.
- Holroyd, K. A. Cognition and desensitization in the group treatment of test anxiety. *Journal of Consulting and Clinical Psychology*, 1976, 44, 991-1001.
- Husek, T. R., & Alexander, S. The effectiveness of the Anxiety Differential in examination stress situations. *Educational and Psychological Measurement*, 1963, 23, 309-318.
- Liebert, R. M., & Morris, L. W. Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, 1967, 20, 975-978.
- Mahoney, M. J. *Cognition and behavior modification*. Cambridge, Mass.: Ballinger, 1974.
- Mallesson, N. Panic and phobia: A possible method of treatment. *Lancet*, 1959, 1, 225-227.
- Mandler, G., & Sarason, S. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, 47, 166-173.
- Meichenbaum, D. H. Cognitive modification of test anxious college students. *Journal of Consulting and Clinical Psychology*, 1972, 39, 370-380.
- Meichenbaum, D. H. *Cognitive behavior modification*. New York: Plenum Press, 1977.
- Meichenbaum, D. H., Gilmore, J. B., & Fedoravich, A. Group insight versus group desensitization in treating speech anxiety. *Journal of Consulting and Clinical Psychology*, 1971, 36, 410-421.
- Morris, L. W., & Liebert, R. M. Effects of anxiety on timed and untimed intelligence tests: Another look. *Journal of Consulting and Clinical Psychology*, 1969, 33, 240-244.
- Osarchuk, M. A comparison of a cognitive, a behavior therapy and a cognitive plus behavior therapy treatment of test anxious college students. (Doctoral dissertation, Adelphi University, 1974). *Dissertation Abstracts International*, 1976, 36, 3619B. (University Microfilms No. 76-1425, 177).
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *The State-Trait Anxiety Inventory (STAI) test manual for form X*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Suinn, R. M. The STABS, a measure of test anxiety for behavior therapy: Normative data. *Behaviour Research and Therapy*, 1969, 7, 335-339.
- Thorpe, G. L. Desensitization, behavior rehearsal, self-instructional training and placebo effects on assertive-refusal behavior. *European Journal of Behavioural Analysis and Modification*, 1975, 1, 30-44.
- Trexler, L. D., & Karst, T. O. Rational-emotive therapy, placebo, and no-treatment effects on public-speaking anxiety. *Journal of Abnormal Psychology*, 1972, 79, 60-67.
- Watson, D., & Friend, R. Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, 1969, 33, 448-457.
- Wilson, G. T., & Davison, G. C. Process of fear reduction in systematic desensitization: Animal studies. *Psychological Bulletin*, 1971, 76, 1-14.
- Wine, J. Test anxiety and direction of attention. *Psychological Bulletin*, 1971, 76, 92-104.
- Wolfe, J. L., & Fodor, I. G. Modifying assertive behavior in women: A comparison of three approaches. *Behavior Therapy*, in press.
- Wonderlic, E. F. *Wonderlic personnel tests*. Northfield, Ill.: Wonderlic, 1959.

Psychological Androgyny and Interpersonal Behavior

Jerry S. Wiggins and Ana Holzmuller
University of British Columbia, Vancouver, Canada

Bem's measure of psychological androgyny was derived from only two relatively desirable dimensions of interpersonal behavior that may, or may not, implicate other less desirable traits that are sex role stereotyped. From an item pool of 1,710 trait-descriptive adjectives, sets of masculinity and femininity scales were assembled that were comparable to "traditional" scales and to those developed by Bem and by Heilbrun. The pool also contained items from eight scales that form an interpersonal circumplex. One hundred eighty-seven college men and women who rated themselves on the 1,710 adjectives were classified as stereotyped, near-stereotyped, or androgynous by Bem's criteria. Bem's measure of psychological androgyny appears to reflect a highly generalizable personological construct that implicates both desirable and undesirable dimensions of interpersonal behavior. Heilbrun's scales are both empirically and conceptually similar to Bem's, and both scale sets differ from traditional masculinity-femininity measures. There is a possibility that androgynous men are more flexible in their interpersonal behavior than androgynous women.

Until very recently, the psychological construct of "masculinity-femininity" has been characterized by a conceptual fuzziness that has permitted unrelated and even contradictory attributes of persons to be viewed as indicants of a monolithic process. Constantine's (1973) critical review of the major psychological tests of masculinity-femininity, developed during the last 40 years, cleared the air for new approaches to an old problem. Bem's (1974) work on the construct of psychological androgyny is one such new approach. Her formulations place a unique emphasis on certain classes of *interpersonal behavior* that are generally considered more

desirable or appropriate for one sex than the other.

Bem (1974) asked male and female judges to rate the desirability of traits for men and women separately. Traits that were rated by both sexes as more desirable for men than for women were called *masculine*, and traits that were rated by both sexes as more desirable for women than for men were called *feminine*. Traits that were rated by both sexes as equally desirable for men and women were called *neutral*. Scales were developed from these item pools and combined to yield a measure of "psychological androgyny." Specifically, the algebraic difference between femininity and masculinity scales was taken to be an index of androgyny. Individuals scoring high on this index are feminine stereotypes, individuals with negative scores are masculine stereotypes, and individuals with scores near zero are androgynous. In a series of studies, Bem (1975; Bem & Lenney, 1976; Bem, Martyna, & Watson, 1976) has marshaled evidence that androgynous persons are flexible in their social behavior and that they can vary their behavior according to situational demands, rather than according to sex role stereotypes.

This research was supported by Canada Council Research Grant S75-0109. We are grateful to Lewis R. Goldberg for making available the large data bank on which our analyses were conducted. Computer programming assistance was graciously provided by Frank Flynn. We would also like to acknowledge the help in data processing provided by Doreen Ridgeway, Barbara J. Marston, and Cecilie Lacey.

Requests for reprints should be sent to Jerry S. Wiggins, Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5.

Other than restricting her inquiry to desirable interpersonal traits, Bem did not specify the manner in which the universe of content of interpersonal behavior was defined. There is now considerable precedent for defining the universe of content of interpersonal behavior with reference to a two-dimensional circumplex of variables of the kind illustrated in Figure 1 (e.g., Becker & Krug, 1964; Benjamin, 1974; Carson, 1969; Foa & Foa, 1974; Leary, 1957; Lorr & McNair, 1963; Rinn, 1965; Schaefer, 1959; Stern, 1970; Swensen, 1973). The system represented in Figure 1 was developed as part of a larger project whose eventual aim is a comprehensive taxonomy of trait-descriptive terms (Wiggins, Note 1). Like the Leary (1957) system of interpersonal behavior on which it is based, the eight variables illustrated in Figure 1 can be decomposed into 16 more narrowly defined variables. In principle, the number of different variables is limited only by the reliability with which respondents can distinguish between closely synonymous words and phrases.

Inspection of the items comprising Bem's Masculinity scale revealed that virtually all of the classifiable items fell within the dominant-ambitious vector of our interpersonal model (e.g., assertive, ambitious, dominant, forceful). Similarly, most of the classifiable

items in Bem's Femininity scale fell within the warm-agreeable vector of the same circumplex (e.g., affectionate, compassionate, sympathetic, tender). Lippa (1977) reported the results of an item analysis of the Bem Sex-Role Inventory (BSRI) in which the items *dominant* and *assertive* were the best markers of the Masculinity scale and the items *compassionate* and *tender* were the best markers of the Femininity scale. Bem (in press) is well aware of the interpersonal content of her scales and uses the terms *instrumental* and *agentic* to describe the Masculinity scale and the terms *expressive* and *communal* to describe the Femininity scale.

Since the two vectors at issue are known to be orthogonal, it is clear why Bem discovered a nonbipolar dimension of masculinity-femininity. But it is also clear that Bem's definition of psychological androgyny is based on only two of the eight major dimensions of interpersonal behavior and that it ignores possible sex role stereotypes related to *undesirable* interpersonal behavior. It could also be the case that the flexibility associated with Bem's definition of psychological androgyny is a more general personality characteristic that subsumes sex role stereotypes. According to the logic of the original Leary (1957) system, the self-actualized individual is one

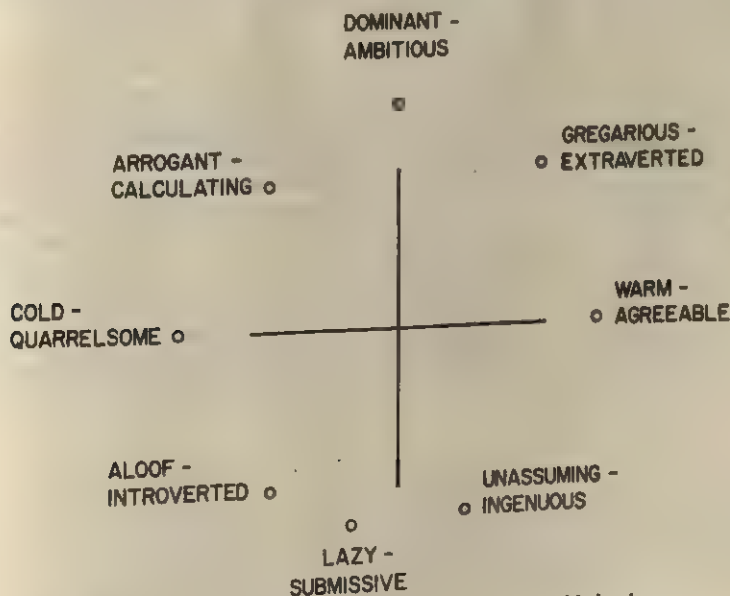


Figure 1. Eight-variable representation of interpersonal behavior.

who is capable of responding in all interpersonal dimensions as the situation demands. Hence, the actualized person is one who has a relatively flat profile of interpersonal traits. Such an individual would, by definition, be androgynous. However, an androgynous person would not necessarily be actualized by Leary's definition.

In the research to be reported here, the self-presentations of masculine-stereotyped, feminine-stereotyped, and androgynous persons are examined with reference to a much broader universe of personological content than has previously been used. Although particular attention was paid to the interpersonal dimensions illustrated in Figure 1, the universe of content sampled was rigorously representative of the entire domain of what Allport and Odbert (1936) called *stable biophysical traits*. In addition, the relationship between bipolar and orthogonal definitions of sex differences was examined in the medium of self-report.

Method

Subjects

The data used in the present analyses were provided by Lewis R. Goldberg. A group of 204 students in an introductory personality class at the University of Oregon rated the self-applicability of 1,710 trait-descriptive adjectives on a 9-place Likert scale ranging from "extremely inaccurate" to "extremely accurate." The students received course credit for this exercise, and they were promised feedback at a later date. The adjectives were partitioned into 10 blocks, and the order in which these blocks were administered was counterbalanced across the sample. The students were allowed to complete the test forms at their own pace over a period of weeks under specific instructions to stop when fatigued or bored. The resultant protocols were checked for carelessness, as indicated by semantic inconsistency, and 17 subjects were eliminated on this basis. The total number of complete usable records was thus 187 (117 females and 70 males).

Item Pool

The pool of items used has historical roots in the work of Allport and Odbert (1936), who identified 4,504 terms in Webster's (1925) *New International Dictionary* that were considered genuine personality traits. Norman (Note 2) expanded this list by consulting Webster's *Third New International Dictionary*,

Unabridged, and, by a variety of procedures, reduced the new list to 2,800 terms. This revised list was, in turn, reduced by Goldberg and Norman to the 1,710 adjectives used in the present study. This list of 1,710 adjectives (Goldberg, Note 3) is thus broadly representative of terms in the English language that describe stable personality traits.

Sex Role Inventory

Forty-three of the 60 items in the BSRI appear in the item pool just described. The majority of items that do not appear are phrases rather than single items. Consequently, single adjectives were identified, in the 1,710 pool, that were closely synonymous with the BSRI phrases (e.g., "decisive" for "makes decisions easily"). Synonymous adjectives were easily found for all but two of the adjectives and phrases in the BSRI: "athletic" on the Masculinity scale and "loves children" on the Femininity scale. Thus, our revised single-adjective sex role inventory consisted of a 19-item Femininity scale, a 19-item Masculinity scale, and a 20-item Neutral scale.

Lacey (Note 4) administered both the original BSRI and our single-adjective sex role inventory to 110 young men and women from the Vancouver, Canada, area. The Femininity, Masculinity, and Neutral scales from the two inventories correlated .92, .97, and .96, respectively, in that sample. Comparable alpha coefficients were also obtained for the two sets of scales. Thus, the measure of psychological androgyny used in the present study is empirically equivalent to Bem's, although it is based on slightly different items.

Interpersonal Circumplex

Sixteen eight-item scales were used to measure the interpersonal dimensions named in Figure 1 (e.g., dominant, ambitious, arrogant, calculating). These 16 scales were scored as octants by combining adjacent variables (e.g., dominant-ambitious, arrogant-calculating). The eight-variable display in Figure 1 is an actual *empirical* plot of the loadings of these eight scales on two hand-rotated principal components extracted from the intercorrelations among the scales in the present sample of 187 subjects. Although these scales were mainly developed on the present sample (Wiggins, Note 1), the same clear circumplex structure has been found in samples of Australian and Canadian students as well (Wiggins & Marston, Note 5).

Masculinity-Femininity Scales

We constructed two relatively pure measures of self-reported "gender," as opposed to sex-related traits, by clustering items that almost necessarily would be answered differently by men and women. A five-item Woman scale (feminine, girlish, lady-

like, unfeminine, womanly) had an alpha coefficient of .96 in the total sample. A five-item Man scale (manly, masculine, unmanly, unmasculine, virile) had an alpha coefficient of .93. The correlation between the two scales ($r = -.91$) approached their estimated reliabilities. Needless to say, these two scales, and the items that comprise them, have non-overlapping distributional properties in male and female samples. The scales are, in effect, caricatures of "traditional" masculinity-femininity scales based on items that empirically discriminate men from women.

Although in some respects, a traditional masculinity-femininity scale (Constantinople, 1973, p. 397), the Masculinity-Femininity scale of Gough and Heilbrun's (1965) Adjective Check List (ACL) attempts "to capitalize on both biological and psychological sex differences" (Heilbrun, 1976, p. 185). The items were selected on the basis of their ability to discriminate between college males identified with masculine fathers and college females identified with feminine mothers (Cosentino & Heilbrun, 1964). Inspection of the 28-item masculine subscale and the 26-item feminine subscale reveals that the scales have highly similar contents to Bem's Masculinity and Femininity scales. The items *masculine* and *feminine* are included in the two scales, as they are in Bem's, but in general the two scales are saturated with dominance and nurturance, respectively. Although originally scored as a single scale (masculine minus feminine), normative and psychometric information is now available for both subscales (Heilbrun, 1976). In the present pool of 1,710 adjectives, all but 2 adjectives ("handsome" and "strong") were available for the masculine subscale and all but one ("praising") for the feminine subscale. The self-applicability of the adjectives was evaluated using a 9-place scale rather than with the usual checklist format.

Experimental Design

In the sample of 187 subjects, we identified groups of stereotyped, near-stereotyped, and androgynous

men and women using Bem's (1974) *t*-ratio criteria.¹ Table 1 shows the percentage of subjects in each of the groups. The 11 "sex-reversed" subjects were not included in the present analysis. Within a 3×2 analysis of variance design, subjects were classified by group (stereotyped, near-stereotyped, androgynous) and by sex (men, women). Scores on the interpersonal adjective scales were entered separately as dependent variables. For each interpersonal variable, this permitted an evaluation of main effects for groups (stereotyped and androgynous groups irrespective of gender), for gender (sex differences in responding irrespective of groups), and for the interaction between these two factors.

Results and Discussion

Sex Roles and the Interpersonal Circumplex

As expected, a main effect for groups was not found for any of the eight interpersonal variables. Stereotyped, near-stereotyped, and androgynous groups did not differ in self-report on the interpersonal variables when the sexes were combined. However, as can be seen from Table 2, there were highly significant sex differences for all but one (aloof-introverted) of the interpersonal variables. Men scored higher on dominant-ambitious, arrogant-calculating, and cold-quarrelsome, whereas women scored higher on lazy-submissive, unassuming-ingenuous, warm-agreeable, and gregarious-extraverted. Sex differences in self-report are not restricted to the relatively desirable vectors of dominance and nurturance; they occur in almost all sectors of the interpersonal circumplex.

To what extent are the sex differences on these interpersonal variables related to the

Table 1
Percentage of Subjects in the Oregon Sample
Classified as Masculine, Feminine, or
Androgynous

Classification	Males ^a	Females ^b
Feminine ($t \geq 2.026$)	3	36
Near feminine ($1.01 < t < 2.026$)	9	28
Androgynous ($-1.01 \leq t \leq +1.01$)	47	33
Near masculine ($-2.026 < t < -1.01$)	30	1
Masculine ($t \leq -2.026$)	11	2

^a $n = 70$.

^b $n = 117$.

¹ This *t* ratio is based on the algebraic difference between femininity and masculinity scores for each subject. There are a number of shortcomings to such a measure (Strahan, 1975), not the least of which is its inability to distinguish androgynous subjects who score high on both masculinity and femininity from "androgynous" subjects who score low on both masculinity and femininity. As a consequence, Bem (1977) now recommends a fourfold classification based on median splits on masculinity and femininity. However, designs such as the present one that analyze interactions between sex role classification and other variables require the additional "near-stereotyped" groups yielded by the difference score method. Within this design, it is still possible to note differences between high-high and low-low androgynous subjects.

Table 2
Analyses of Variance of Interpersonal Variables

Variable	Sex (A)		Group (B)		A × B	
	F	p	F	p	F	p
Dominant-ambitious	21.3	.00001	.3	ns	30.6	.00000
Arrogant-calculating	24.6	.00000	.3	ns	3.4	.04
Cold-quarrelsome	45.4	.00000	.3	ns	2.3	ns
Aloof-introverted	2.3	ns	1.9	ns	7.5	.0008
Lazy-submissive	15.2	.0001	2.3	ns	14.6	.00000
Unassuming-ingenuous	39.1	.00000	1.3	ns	9.2	.0002
Warm-agreeable	33.2	.00000	1.3	ns	5.9	.003
Gregarious-extraverted	13.8	.0003	1.7	ns	2.6	ns

stereotyped, near-stereotyped, and androgynous classifications of the sex role inventory? From the final column of Table 2, it can be seen that significant interactions between gender and sex role classification occurred on six of the eight interpersonal variables. The pattern of these interactions is consistent with Bem's concept of androgyny for all but one (again, aloof-introverted) of the six interpersonal variables. The greatest difference in self-report occurred between male and female stereotypes, the next greatest between male and female near stereotypes, and the least difference between male and female androgynous subjects. Despite substantial overall sex differences on arrogant-calculating, lazy-submissive, unassuming-ingenuous, and warm-agreeable, there were no statistically reliable differences between androgynous men and women on these variables.² Although the difference between androgynous men and women on dominant-ambitious was statistically significant, it involved a "crossover" effect that is illustrated in Figure 2. Androgynous women presented themselves as significantly *more* dominant-ambitious than androgynous men. A similar crossover occurred on lazy-submissive, although the difference between androgynous men and women was not significant.

These results suggest that Bem's measure of androgyny reflects a highly generalizable personological construct. Persons classified as androgynous by the sex role inventory are nonstereotypic, not only in the realms of dominant-ambitious and warm-agreeable behavior but in the realms of arrogant-calculating, lazy-submissive, and unassuming-ingenu-

ous behaviors as well. Persons classified as sex role stereotyped by the Bem inventory are stereotyped on the same five dimensions of interpersonal behavior.

The data suggest that the undesirable dimension of cold-quarrelsome behavior is not the functional opposite of the desirable warm-agreeable dimension, with respect to sex role stereotypy. The trend of cold-quarrelsome is clearly for the largest sex differences to occur in the stereotyped groups and the smallest in the androgynous groups. However, the interaction of sex and group classification was not significant, and there was no trend toward a crossover effect as there was with warm-agreeable. In general, none of the subjects presented themselves as particularly cold and quarrelsome, although men were more willing to do so than women.

Striking exceptions to the foregoing trends occurred with reference to aloof-introverted and gregarious-extraverted, which represent poles of the ubiquitous "introversion-extraversion" dimension of personality research. On both variables the differences between androgynous men and women were significant, whereas those between stereotyped men and women were not. The results of the analysis of aloof-introverted are particularly anomalous. As can be seen from Figure 3, the two stereotyped groups are indistinguishable, as are the two near-stereotyped groups. However, androgynous women presented them-

² All post hoc comparisons between sex role groups were evaluated by the Newman-Keuls procedure with alpha set at $p \leq .05$.

selves as significantly less introverted than androgynous men, and, in fact, significantly less introverted than all groups. Thus, there is a significant interaction between sex and group classification in the absence of significant main effects for the latter two factors. A similar trend emerged from the analysis of gregarious-extraverted, although no interaction or crossover trend was evident. All female groups reported more extraversion than their male counterparts, although the only significant difference occurred between androgynous women and men.

It is tempting to speculate on the reasons for the lack of generalizability of Bem's index of psychological androgyny to the domains of introverted and extraverted behaviors. It appears that introversion is not a sex stereotyped role, as evidenced by the lack of overall gender differences, although androgynous women clearly see themselves as less introverted than their male counterparts. Although there are significant gender differences in self-reported extraversion, the lack of an interaction with sex role classification raises the possibility that this relatively desirable variable is not sex stereotyped either. If one thinks of the gregarious-extraverted social role as a blend of dominance and nurturance, then it is understandable that both men and women could assume it without violating stereotypes.

Sex Roles and Trait-Descriptive Adjectives

To what extent do the sex role groups differ among themselves on variables other than those of the interpersonal domain? The 1,710 adjectives to which all subjects responded embrace a variety of personal characteristics other than the strictly interpersonal. Although a final taxonomy of this item pool has not yet been achieved, a preliminary set of distinctions has been made among the domains of *interpersonal traits* (dominant, warm), *temperamental traits* (animated, high-strung), *character traits* (moral, unprincipled), *material traits* (miserly, materialistic), *attitudes* (prejudiced, progressive), *mental predicates* (analytic, intelligent), and *social roles* (aristocratic, cliquish) (Wiggins, Note 1). Differ-

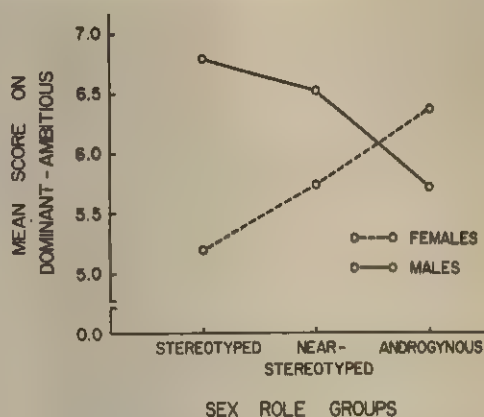


Figure 2. Mean scores of male and female sex role groups on dominant-ambitious.

ences in self-report between stereotyped, near-stereotyped, and androgynous groups were examined in all of these domains by comparing the group means on all 1,710 adjectives. Even with highly stringent significance levels, this procedure produced an *embarras de richesses* that is not easily summarized nor confidently interpreted. Nevertheless, several trends were apparent from these data that are worth noting, if only as guides for future investigations.

With remarkably few exceptions, the adjectives that differentiated sex role groups from each other were *interpersonal* in nature and could be easily classified within the 8-variable system used in the main analysis or

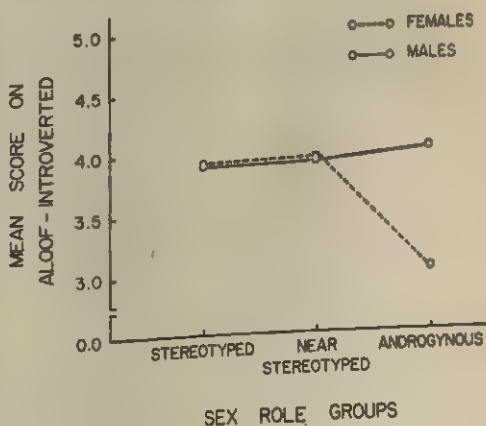


Figure 3. Mean scores of male and female sex role groups on aloof-introverted.

within the more refined 16-variable system. In virtually all cases, interpersonal adjectives that differentiated two sex role groups did so in a manner that was consistent with the findings of the analyses of variance reported earlier. In comparing androgynous with masculine males, the former presented themselves as warm and submissive and the latter as dominant and cold. In comparison with feminine females who described themselves as meek and submissive, androgynous females described themselves as dominant and extraverted. The comparisons of androgynous and stereotyped groups on the 1,710 adjectives mainly confirmed the earlier analyses with many more interpersonal adjectives.

The adjectives that differentiated androgynous men from androgynous women served to confirm the earlier analysis and to extend it somewhat into the domain of temperament. The 51 adjectives listed in Table 3 reliably distinguished ($p < .001$) the two groups. On the item *aggressive*, for example, the mean item response for androgynous females was 6.54 and the mean item response for androgynous males was 4.82. The item *antisocial* was answered as significantly more self-descriptive by androgynous males than by androgynous females. The starred items are significant under an ultraconservative criterion ($p = .05/1,710 = .000029$). In comparison with the androgynous male, the androgynous female presents herself as extraverted, warm, excitable, emotional, aggressive, and vivacious. In

contrast, the androgynous male is introverted, cold, calm, unemotional, passive, and undramatic. In this sample of subjects, at least, males and females achieved psychological androgyny by different routes. Although both androgynous groups clearly differed from stereotyped groups of the same gender, they also differed from each other in ways that both support and contradict sex role stereotypes.

The present data also permitted a test of the hypothesis that subjects who are classified as androgynous on the basis of high and equal scores on femininity and masculinity (high-high) differ from subjects who achieve that classification by low and equal scores (low-low), primarily with respect to *self-esteem* (Spence, Helmreich, & Stapp, 1975). Like Bem (1977), we were unable to find many examples of truly "low-low" androgynous subjects in our sample. Nevertheless, we were able to gain some information on this issue by contrasting the upper and lower 25% of each androgynous group. Table 4 displays the adjectives that differentiated high-high and low-low androgynous subjects under this definition.

From among the considerable number of adjectives that differentiated high-high from low-low androgynous females, we have reported only those that did so at the conservative significance level. Overall it is clear that the high-highs attributed positive characteristics to themselves, whereas the low-lows

Table 3

Items that Discriminated Androgynous Females from Androgynous Males

Females (39)		Males (33)	
Aggressive*	Naive	Antisocial	Reserved
Animated	Open-hearted	Bashful	Shy*
Blunt	Overexcitable	Calm	Silent
Bossy*	Overtalkative*	Emotionless	Soft-spoken
Dainty*	Peppy	Feelingless	Taciturn
Direct	Perky	Hard-hearted	Tough
Emotional	Rambunctious	Indirect	Unaggressive
Excitable	Sollicitous	Iron-hearted	Undramatic
Extraverted*	Talkative	Meek	Unemotional
Fickle	Uncalculating	Overquiet	Unexcitable
Frightenable	Unshy	Passive	Unfeeling
Genial	Vivacious	Quiet*	Unromantic
Gullible*	Vocal	Quiet-spoken	

* $p < .000029$. All other items significant at $p < .001$.

Table 4

Items that Discriminated High-High from Low-Low Androgynous Groups

Females		Males	
High-high ^a	Low-low ^a	High-high ^b	Low-low ^b
Affectionate	Disobliging	Assertive	Aimless
Certain	Feelingless	Compassionate	Compassionless
Courteous	Heartless	Comradely	
Friendly	Hostile	Forceful	
Genial	Ill-tempered	Forthright	
Outgoing	Joyless	Heroic	
Polished	Peevish	Magnetic	
Thoughtful	Pessimistic	Penetrative	
	Self-doubting	Self-respecting	
	Spineless	Self-seeking	
	Unassured	Steadfast	
	Unpleasable		

Note. Differences between female groups significant at $p < .000029$. Differences between male groups significant at $p < .001$.

^a $n = 10$.

^b $n = 8$.

attributed negative characteristics to themselves. More germane to the hypothesis, the high-highs described themselves on the rating scale as more *certain* ($M = 6.90$) than did the low-lows ($M = 3.90$). Similarly, the low-lows described themselves as more *self-doubting* (6.30 vs. 3.00) and as more *unassured* (5.10 vs. 2.00).

There were fewer adjectives that differentiated the male groups, and consequently a less stringent significance level ($p < .001$) was used to detect trends. As was true of the females, the high-high males attributed highly positive characteristics to themselves in contrast to the low-low males. The high-highs presented themselves as more *self-respecting* (8.00) than did the low-lows (5.75). The low-lows described themselves as more *aimless* (4.62) than did the high-highs (1.87). In general, the data for both men and women suggest that high-high androgynous subjects have considerably more self-esteem than do low-low subjects. This trend is especially interesting in the present sample in which high-highs and low-lows differed only slightly from each other in their masculinity and femininity scores.

The fact that high-high and low-low androgynous subjects differed in self-esteem could restrict the generalizability of the conclusions drawn concerning differences between

androgynous males and androgynous females (Table 3). Consequently, differences in self-report between androgynous males and androgynous females were reanalyzed excluding the lower 25% of both groups, a liberal definition of low-lows. None of the findings in Table 3 change when low-low androgynous subjects are excluded. Where differences between high-high and low-low groups occurred, they served only to accentuate the trends reported in Table 3.

Masculinity-Femininity Scales and Interpersonal Behavior

Three different sets of masculinity and femininity scales were scored in the present sample: (a) Bem's Masculinity and Femininity scales from the BSRI; (b) Heilbrun's Masculinity and Femininity scales from the ACL; and (c) the five-item Man and Woman scales that were constructed as markers of traditional masculinity-femininity scales. The intercorrelations among these scales, and their reliabilities, are presented in Table 5. The BSRI and ACL scales are of comparable homogeneity, and the Man and Woman scales are, as expected, highly homogeneous despite the small number of items involved.

The correlations between comparable BSRI and ACL scales suggest that the two sets of

Table 5

Intercorrelations Among Femininity, Masculinity, and Femininity Minus Masculinity Scores

Variable	1	2	3	4	5	6	7	8
1. BSRI-F	(.718)							
2. BSRI-M	-.127	(.830)						
3. BSRI-D	.678	-.816						
4. ACL-F	.731	-.121	.516	(.771)				
5. ACL-M	-.257	.873	-.800	-.217	(.831)			
6. ACL-D	.599	-.686	.857	.724	-.830			
7. Woman	.641	-.354	.636	.627	-.409	.647	(.963)	
8. Man	-.562	.382	-.611	-.578	.423	-.629	-.913	(.932)
9. WM-D	.616	-.376	.638	.617	-.425	.652	.980	-.976

Note. BSRI-F = Bem Sex-Role Inventory Femininity scale; BSRI-M = Bem Sex-Role Inventory Masculinity scale; BSRI-D = difference between the two scales; ACL-F = Adjective Check List Femininity scale; ACL-M = Adjective Check List Masculinity scale; ACL-D = difference between the two scales; WM-D = difference between the Woman and Man scales. For $N = 187$, an r of .187 is significant at the .01 level. The reliabilities (Cronbach's alpha) of the masculinity and femininity scales appear in parentheses. The correlation coefficients in italics indicate the degree of bipolarity that exists between masculinity and femininity scales in the three sets.

measures are closely related. The masculinity measures are correlated .873, the femininity measures are correlated .731, and the difference measures (femininity minus masculinity) are correlated .857. Both the BSRI and the ACL scale sets are substantially correlated with the traditional masculinity-femininity scales. The ACL Masculinity scale was more highly correlated with the Man scale than was the BSRI Masculinity scale, but the BSRI Femininity scale was more highly correlated with the Woman scale than was the ACL Femininity scale. The BSRI difference score (femininity-masculinity) was correlated .628 with woman minus man, and the ACL difference score was correlated .564 with the same index. Although the BSRI and the ACL scale sets are conceptually distinct from the traditional masculinity-femininity measures, they share a considerable amount of variance in common. The extent of these relationships may be exaggerated by the use of a common method of measurement (9-place Likert scales) in the present study, but it is expected that the scales would also be highly correlated in their original formats.

The traditional Man and Woman scales are strongly bipolar ($-.913$), the ACL scales are only slightly related ($-.217$), and the BSRI scales are the most independent ($-.127$). The BSRI and ACL scales also differ from the traditional scales in the extent to which mascu-

linity and femininity determine the difference score (femininity minus masculinity). For the traditional scales, the correlation between man and woman minus man ($-.976$) was almost identical to the correlation between woman and woman minus man (.980). For the BSRI scales, however, the correlation between masculinity and the difference score ($-.816$) was higher than the correlation between femininity and the difference score (.678). The same is true for the ACL scales. This suggests that in the present sample at least the greatest contribution to variance on the androgyny index comes from the Masculinity scale.

The conceptual similarity of the ACL scales to the BSRI scales is most evident from the pattern of correlations of these two scale sets with the interpersonal variables used in the present study. From Table 6 it can be seen that the masculinity and femininity scales from the BSRI and ACL behave as variables that are closely related to the circumplex positions of dominant-ambitious and warm-agreeable, respectively. The masculinity scales have their highest positive correlations with dominant-ambitious and their highest negative correlations with the opposite vector of lazy-submissive. The femininity scales have their highest positive correlations with warm-agreeable and their highest negative correlations with the opposite vector of cold-quarrelsome. This orthogonal pattern is also clearly evi-

dent in the correlation of the androgyny indices (feminine minus masculine) with the interpersonal variables. From Table 6 it can be seen that the ACL androgyny index has its two highest positive correlations with lazy-submissive and warm-agreeable and its two highest negative correlations with dominant-ambitious and cold-quarrelsome. The same pattern is evident for the BSRI androgyny index, with the slight exception that the negative correlation with arrogant-calculating exceeds the negative correlation with cold-quarrelsome. Aside from this slight exception, the patterns of correlations of the BSRI and ACL scale sets with the interpersonal circumplex variables are virtually identical.

The pattern of correlations of the traditional Man and Woman scales with the interpersonal variables is quite different. The Man, Woman, and Woman minus Man scales all have their highest correlations with the single bipolar dimension of cold-quarrelsome versus warm-agreeable. Masculinity is associated with hostility and femininity with warmth. The apparent association between masculinity and hostility is especially interesting, since none of the items in the Man or Woman scales implicate interpersonal variables.

Psychological Androgyny or Interpersonal Flexibility?

According to Bem, the androgynous person is both dominant and nurturant and thus can vary his or her behavior to meet situational demands. Stereotyped males would, presumably, emphasize dominance and suppress nurturance, and stereotyped females would emphasize nurturance and suppress dominance. Earlier in this article we raised the possibility that the flexibility of androgynous persons may be part of a broader pattern of flexibility that is expressed in all or most dimensions of interpersonal behavior. By this reasoning, the androgynous person's profile of interpersonal variables would be relatively flat, and the stereotyped person's profile would be both positively and negatively spiked on variables that are highly sex stereotyped.

The profile variability of sex role groups

Table 6
Correlations Between Masculinity-Femininity Scale Sets and Interpersonal Variables

Scale	Dominant-ambitious	Arrogant-calculating	Cold-quarrelsome	Alloof-introverted	Lazy-submissive	Unassuming-ingenuous	Warm-agreeable	Gregarious-extraverted
BSRI-F	-.050	-.266	-.609	-.175	.168	.283	.750	.388
BSRI-M	.797	.512	.204	-.251	-.600	-.291	-.073	.374
BSRI-D	-.620	-.534	-.506	.084	.540	.381	.492	-.051
ACL-F	-.072	-.124	-.521	-.281	.206	.141	.731	.478
ACL-M	.763	.678	.376	-.126	-.566	-.413	-.219	.241
ACL-D	-.580	-.550	-.563	-.071	.517	.372	.572	.103
Woman	-.134	-.316	-.453	-.148	.186	.219	.436	.257
Man	.173	.316	.431	-.178	-.186	-.188	-.379	-.240
WM-D	-.156	-.323	-.452	-.166	.190	.208	.418	.255

Note. BSRI-F = Bem Sex-Role Inventory Femininity scale; BSRI-M = Bem Sex-Role Inventory Masculinity scale; BSRI-D = difference between the two scales; ACL-F = Adjective Check List Femininity scale; ACL-M = Adjective Check List Masculinity scale; ACL-D = difference between the two scales; WM-D = difference between the Woman and Man scales. For $N = 187$, an r of .187 is significant at the .01 level.

was examined with reference to scores on the full set of 16 interpersonal variables. Sixteen variables were used to ensure an adequate number of observations for profile analysis. Means and standard deviations were computed separately in the total group of men and in the total group of women for each of the 16 variables. All subjects' raw scores were then converted to standard scores based on either male or female norms. Mean standard scores were computed for male and female stereotyped, near-stereotyped, and androgynous groups. This resulted in a mean profile of 16 interpersonal variables, for each sex role group, that had been standardized with reference to the total same-sex sample. The variance of a group's profile of mean standard scores was used as an index of profile variability for that group. Differences in profile variability between sex role groups were assessed by a variance ratio (F) with 15, 15 degrees of freedom.

The hypothesis that androgynous persons would have relatively flat profiles of interpersonal variables received considerable support for male subjects. Androgynous males had a flat profile of scores that were close to the means for the total group. In contrast, the profile of scale scores for stereotyped males was much more variable, with a high spike on dominant and low spikes on warm and submissive. The interpersonal profile of stereotyped males was significantly ($p < .001$) more variable than those of near-stereotyped males and androgynous males. The profile of near-stereotyped males was slightly ($p < .02$) more variable than that of androgynous males.

The hypothesis that androgynous persons would have relatively flat profiles of interpersonal variables was strikingly disconfirmed for female subjects. The index of profile variability was highest for androgynous females and next highest for stereotyped females; the slight difference between the two groups was not statistically significant. Near-stereotyped females were least variable, and their relatively flat profile was significantly ($p < .001$) less variable than those of androgynous and stereotyped females.

Stereotyped females had a profile that was

high on introverted, submissive, and unassuming, and was low on dominant, extraverted, and arrogant. The profile of androgynous females was a *mirror image* of this pattern. The correlation between the profile of stereotyped females and that of androgynous females was $-.96$ ($p < .001$). Rather than differing from stereotyped females in profile variability, androgynous females were equally variable, but in ways that run counter to sex role stereotypes. Androgynous males differ from stereotyped males in both the mean level and variability of their profile of interpersonal variables. In contrast, androgynous females are, interpersonally, the *opposite* of their stereotyped counterparts. The profile of androgynous females was significantly ($p < .001$) more variable than that of androgynous males, and it is again clear that androgyny does not reflect a uniform set of dispositions in men and women. To the extent that a flat profile of variables is an index of interpersonal flexibility, androgynous women would be characterized as "inflexibly androgynous," whereas androgynous men would be characterized as "flexibly androgynous."

To investigate the possibility that the observed patterns of profile variability may have been affected by the presence of low-low androgynous subjects, the data were analyzed with the lower 25% of both androgynous groups excluded. This had the effect of making the profile of androgynous males slightly less variable and the profile of androgynous females slightly more variable. All of the major trends previously noted remained the same.

Conclusions

Despite the fact that it was derived from only two relatively desirable dimensions of interpersonal behavior, Bem's measure of psychological androgyny appears to reflect a highly generalizable personological construct. Persons classified as androgynous or as stereotyped by the BSRI will tend to be androgynous or stereotyped on five of the eight major dimensions of interpersonal behavior. Although there was a trend for this to be true of the highly undesirable cold-quarrelsome

dimension, it was not a significant one. A genuine exception to the generalizability of BSRI sex role classifications occurred with respect to the dimensions of introversion and extraversion. Women were more extraverted than men in all sex role groups, with the greatest differences occurring between androgynous groups. Although stereotyped groups did not differ on introversion, androgynous women were clearly less introverted than androgynous men. The possibility that the dimensions of introversion and extraversion are not sex role stereotyped is worth investigating in other samples of subjects.

Sex role groups defined by the BSRI differed from each other in self-report primarily on *interpersonal adjectives* as opposed to adjectives denoting temperamental, characterological, material, attitudinal, or intellectual characteristics. Androgynous groups differed from stereotyped groups of the same gender, but they also differed from each other in ways that both support and contradict sex role stereotypes. Androgynous men and women who scored high on both masculinity and femininity scales differed from their same-gender androgynous counterparts who scored low on both scales. The high-high androgynous subjects characterized themselves as having greater self-esteem than did low-low androgynous subjects (Spence et al., 1975).

The Masculinity and Femininity scales from Gough and Heilbrun's (1965) ACL are much more closely related to the corresponding Bem scales than they are to traditional masculinity-femininity measures. In addition to being highly correlated with the Bem scales, the ACL Masculinity and Femininity scales exhibit the same pattern of correlations with interpersonal variables. Thus, it seems appropriate to characterize subjects who score high on both scales as psychologically androgynous (Heilbrun, 1976).

When the 16-variable interpersonal profiles of sex role groups were compared on profile variability, sex differences were again found between androgynous groups. Androgynous males had a relatively flat profile of interpersonal variables that stood in contrast to the profile of stereotyped males that spiked on sex role stereotyped variables. Androgynous

females had a variable profile that was a mirror image of the sex role stereotyped profile for stereotyped females. Such a finding does not necessarily warrant the conclusion that androgynous males are more "flexible" in their interpersonal behavior than androgynous females. The greater profile variability of androgynous females may reflect a more differentiated self-perception on their part. Whether or not such differences in self-report are reflected in behavior is, of course, still an open, empirical question.

Reference Notes

1. Wiggins, J. S. *A psychological taxonomy of trait-descriptive terms: I. The interpersonal domain*. Paper presented at the meeting of the Society of Multivariate Experimental Psychology, Gleneden Beach, Oregon, November 1975.
2. Norman, W. T. *2800 personality trait descriptors: Normative operating characteristics for a university population*. Ann Arbor: Department of Psychology, University of Michigan, 1947. (Mimeo)
3. Goldberg, L. R. *Some normative characteristics of the 1,710 trait descriptive adjectives*. Eugene: Oregon Research Institute, 1974. (Mimeo)
4. Lacey, C. *Psychological androgyny and cognitive complexity*. Unpublished undergraduate honor's thesis, University of British Columbia, 1976.
5. Wiggins, J. S., & Marston, B. J. *An index for evaluating circumplexity in two-dimensional real-space models*. Unpublished manuscript, University of British Columbia, 1976.

References

- Allport, G. W., & Odbert, H. S. Trait names: A psycho-lexical study. *Psychological Monographs*, 1936, 47(1, Whole No. 211).
- Becker, W. C., & Krug, R. S. A circumplex model for social behavior in children. *Child Development*, 1964, 35, 371-396.
- Bem, S. L. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1974, 42, 155-162.
- Bem, S. L. Sex role adaptability: One consequence of psychological androgyny. *Journal of Personality and Social Psychology*, 1975, 31, 634-643.
- Bem, S. L. On the utility of alternative procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1977, 45, 196-205.
- Bem, S. L. Beyond androgyny: Some presumptuous prescriptions for a liberated sexual identity. In J. Sherman & F. Denmark (Eds.), *Psychology of*

- women: *Future directions of research*. New York: Psychological Dimensions, in press.
- Bem, S. L., & Lenney, E. Sex typing and the avoidance of cross-sex behavior. *Journal of Personality and Social Psychology*, 1976, 33, 48-54.
- Bem, S. L., Martyna, W., & Watson, C. Sex typing and androgyny: Further explorations of the expressive domain. *Journal of Personality and Social Psychology*, 1976, 34, 1016-1023.
- Benjamin, L. S. Structural analysis of social behavior. *Psychological Review*, 1974, 81, 392-425.
- Carson, R. C. *Interaction concepts of personality*. Chicago: Aldine, 1969.
- Constantinople, A. Masculinity-femininity: An exception to a famous dictum? *Psychological Bulletin*, 1973, 80, 389-407.
- Consentino, F., & Heilbrun, A. B. Anxiety correlates of sex-role identity in college students. *Psychological Reports*, 1964, 14, 729-730.
- Foa, U. G., & Foa, E. B. *Societal structures of the mind*. Springfield, Ill.: Charles C Thomas, 1974.
- Gough, H. G., & Heilbrun, A. B. *Manual for the Adjective Check List and the Need scales for the ACL*. Palo Alto, Calif.: Consulting Psychologists Press, 1965.
- Heilbrun, A. B. Measurement of masculine and feminine sex role identities as independent dimensions. *Journal of Consulting and Clinical Psychology*, 1976, 44, 183-190.
- Leary, T. *Interpersonal diagnosis of personality*. New York: Ronald Press, 1957.
- Lippa, R. The naive perception of masculinity-femininity on the basis of expressive cues. *Journal of Research in Personality*, 1977, 11, 21-37.
- Lorr, M., & McNair, D. M. An interpersonal behavior circle. *Journal of Abnormal and Social Psychology*, 1963, 67, 68-75.
- Rinn, J. L. Structure of phenomenal domains. *Psychological Review*, 1965, 72, 445-466.
- Schaefer, E. S. A circumplex model for maternal behavior. *Journal of Abnormal and Social Psychology*, 1959, 59, 226-235.
- Spence, J. T., Helmreich, R., & Stapp, J. Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 1975, 32, 29-39.
- Stern, G. G. *People in context: Measuring person-environment congruence in education and industry*. New York: Wiley, 1970.
- Strahan, R. F. Remarks on Bem's measurement of psychological androgyny: Alternative methods and a supplementary analysis. *Journal of Consulting and Clinical Psychology*, 1975, 43, 568-571.
- Swensen, C. H. *Introduction to interpersonal relations*. Glenview, Ill.: Scott, Foresman, 1973.
- Webster's new international dictionary* (2nd ed., unabridged). Springfield, Mass.: G. & C. Merriam, 1924.

Received September 30, 1977 ■

Garfield Appointed Editor, 1979-1984

The Publications and Communications Board of the American Psychological Association announces the appointment of Sol L. Garfield as Editor of the *Journal of Consulting and Clinical Psychology* for the years 1979-1984. As of January 1, 1978, manuscripts should be directed to the Editor-elect:

Sol L. Garfield
Department of Psychology
Washington University
St. Louis, Missouri 63130

Comparing the Empirical Validity of the Standard Form with Two Abbreviated MMPIs

Charles S. Newmark and David R. Ziff
University of North Carolina School of Medicine

Alfred J. Finch, Jr.
Virginia Treatment Center for Children
Richmond

Philip C. Kendall
Virginia Commonwealth University

The standard form Minnesota Multiphasic Personality Inventory (MMPI) and two abbreviated forms, the MMPI-168 and the Faschingbauer Abbreviated MMPI (FAM), were compared with direct measures of psychopathology obtained from the Brief Psychiatric Rating Scale (BPRS) with psychiatric inpatients. Each patient was interviewed using the Mental Status Schedule by one rater while another rater observed this initial diagnostic interview behind a one-way mirror. Thus, each patient was rated on the BPRS by two raters to assess interrater reliability. Since MMPI scales contain more than one interpretative factor, these scales were correlated with the means of more than one BPRS symptom using multiple correlation coefficients. The multiple correlation coefficients between the BPRS ratings and the corresponding MMPI and abbreviated-form scales were significantly high and comparable. Only on *Pd* for females did a significant difference occur, with the FAM correlation being significantly higher. These findings suggest that these abbreviated forms are an accurate substitute for the standard-form MMPI in predicting objective measures of psychopathology.

The recent development of several abbreviated forms of the Minnesota Multiphasic Personality Inventory (MMPI) has precipitated a plethora of investigations assessing the practical utility of these instruments. The majority of these studies focused solely on comparisons with the standard MMPI of group mean data and individual profile pairs concerning validity, high points, general elevations, code-type correspondence, configural analyses, and application of various MMPI-derived diagnostic rules. Of the six abbreviated MMPIs now available for clinical use, namely Kincannon's (1968) Mini-Mult, Dean's (1972) Midi-Mult, Hugo's (1971) Short Form, Faschingbauer's (1973) Abbre-

viated MMPI (FAM), Overall and Gomez-Mont's (1974) MMPI-168, and Spera and Robertson's (Note 1) Maxi-Mult, only the FAM and the MMPI-168 have been demonstrated to predict accurately both individual and group mean data with both normal and psychiatric samples (Newmark, Boas, & Messervy, 1974; Newmark, Cook, Clarke, & Faschingbauer, 1973; Newmark, Galen, & Gold, 1975; Newmark & Glenn, 1974; Newmark, Newmark, & Cook, 1975; Newmark, Newmark, & Faschingbauer, 1974; Newmark, Owen, Newmark, Cook, & Faschingbauer, 1975; Newmark & Raft, 1976).

None of these investigations, however, has dealt with two crucial issues. First, Is the interpretation obtained from an abbreviated form comparable to that obtained from the standard form? If not, regardless of whether individual and group mean data are comparable, the abbreviated form should not be used. Therefore, the comparative interpretive efficacy of the FAM and the standard MMPI

Requests for reprints should be sent to Charles S. Newmark, Department of Psychiatry, University of North Carolina Medical School, Chapel Hill, North Carolina 27514.

Philip C. Kendall is now at the University of Minnesota.

was assessed with a sample of psychiatric inpatients (Newmark, Conger, & Faschingbauer, 1976). Psychiatric residents evaluated the accuracy of the interpretation of both test forms for each of their patients. Although significantly higher mean ratings resulted for the standard MMPI, when data were collected across sexes, the quantity of loss was significantly lower than expected. That is, the FAM functioned as an instrument 80% as long as the standard MMPI, even though it contains only 30% of the items.

Newmark, Falk, and Finch (1976) also compared the interpretive accuracy of the standard MMPI and three abbreviated forms with a sample of psychiatric inpatients. Although significantly higher mean ratings resulted from the standard MMPI when compared with either the FAM or Hugo's Short Form, comparable ratings by psychiatric teams were obtained when comparing the MMPI-168 and the standard-form MMPI interpretations. Two major limitations of these latter two studies include the use of psychiatric judges as the validity criterion and the failure to have the interpretations done by computer to reduce the error factor introduced by clinical judgment.

A second and more crucial issue concerns the empirical validity of the standard and abbreviated forms when compared with direct measures of psychopathology. In previous investigations the utility of the abbreviated form has been based on concurrent validity with the standard form of the MMPI. Thus, though the FAM may predict the standard-form MMPI more precisely than the MMPI-168 does, the MMPI-168 might predict psychopathology more accurately than either the FAM or standard-form MMPI.

The present investigation attempted to compare the empirical validity of two abbreviated forms of the MMPI and the standard form with direct measures of psychopathology using psychiatric inpatients. This investigation seems necessary, since there is no reason to use an alternative version of a psychometric instrument, whether longer or shorter, unless it possesses greater empirical validity than the form it replaces or unless it is significantly less expensive to administer with

no loss in empirical validity. Rand (1976) and Overall, Higgins, and De Schweinitz (1976) independently concluded that using external criteria is a vital prerequisite for evaluating abbreviated forms of the MMPI.

Method

Subjects

The subjects were 158 male and 217 female consecutive admissions to either a private or a university psychiatric inpatient facility. For numerous reasons, primarily invalid profiles, lack of cooperation, confusion, limited intellectual ability, and cerebral dysfunction, 38 males and 47 females were eliminated from this investigation. The resultant 120 males and 170 females were between the ages of 17 and 65 years ($M = 34.7$), and their level of education ranged from 3 to 22 years ($M = 10.2$). First admissions to a psychiatric hospital comprised 71% of the sample. Although there were no significant educational differences as a function of sex, the female patients were significantly older ($p \leq .05$).

Apparatus

The Brief Psychiatric Rating Scale (extended version). The rating scale that was used to record clinical observations in this investigation was a 42-item instrument devised by Pokorny, in collaboration with Overall and others, for use in a survey of the Texas state hospital population (Pokorny & Overall, 1970). It consists of the widely used 18-item Brief Psychiatric Rating Scale (BPRS; Overall & Gorham, 1962) plus 24 additional rating constructs from various syndrome scales chosen to span the range of lesser psychopathologies. Each symptom construct in the BPRS, and in the extended version used here, is rated on a 7-point scale of severity. Approximately half of the BPRS ratings are based on observations of the patients, their modes of communication, and the organization of their mental processes without regard to specific content. Information regarding the development, scoring procedure, reliability, and validity of this instrument can be found elsewhere (Overall & Gorham, 1962; Overall & Klett, 1972).

The BPRS was chosen as the criterion variable in this investigation, since it has been used fruitfully for the isolation and identification of psychopathological syndromes and categories in numerous studies (Auerbach & Ewing, 1964; Overall, 1974; Overall, Hollister, Johnson, & Pennington, 1966; Overall, Hollister, & Pichot, 1967; Pokorny & Overall, 1970; Steer, 1974) and because the practical utility of rating scales as objective measures of psychopathology has been extensively documented (Lorr, 1954; Norton, 1967; Schreier, Reznikoff, & Glueck, 1962; Wittenborn, 1972). Furthermore, the BPRS has symptom labels similar to those found on the

MMPI scales. The development of rating scales in an attempt to standardize, fractionate, and objectify the process of clinical decision making has been among the most significant achievement in psychodiagnosis during the past decade (Goldberg, 1974). That behavioral rating scales have largely supplanted psychological tests as criterion measures has been discussed recently by Cleveland (1976).

The Mental Status Schedule. To complete the BPRS, Overall and Gorham (1962) advocated the use of an 18-minute generally nondirective interview. However, this interview proved impractical in the present investigation due to its brevity and lack of structure and because the usefulness of rating scales has been shown to be limited when there is significant variability in the interview procedures on which the ratings are based (Spitzer, Fleiss, Burdock, & Hardesty, 1964). Therefore, the Mental Status Schedule (MSS), which was developed by Spitzer, Burdock, and Hardesty (1964) to provide a standardized interview to assess the major dimensions of the mental status in which the content and order of questions are fixed, was used in this investigation.

The MSS contains an interview schedule and a matching inventory of 248 dichotomous items descriptive of small units of psychopathological behavior that the interviewer evaluates as true or false. The inventory was not used in this study. The interview schedule is a series of 82 questions arranged in a definite sequence to provide a natural progression of topics that cover a wide range of psychopathology that the interviewer uses to elicit information from the patient. Fifty-one supplementary questions were provided to clarify or probe the areas in which the patient's responses seemed incomplete. The schedule provided alternative phraseology for many questions so that the examiner could ask the question in the form and tense most appropriate to the circumstances. Thus, even though standardized, the procedure has enough flexibility so that when properly administered it seems like a typical clinical interview, allowing good rapport between interviewer and patient (Spitzer, Fleiss, Kernohan, Lee, & Baldwin, 1965). The items on the inventory were developed by surveying standard psychiatric texts and by interviewing several hundred psychiatric patients with preliminary forms of the MSS to obtain a comprehensive coverage of overt signs of psychopathology. Efforts were made to word the questions so that even inpatients with minimal education or psychological insight could comprehend them. The length of the interview varied from 20 to 60 minutes depending on the patient's verbal productivity, amount of psychopathology present, and cooperation. A detailed description of the MSS as well as information bearing on the reliability, validity, and administration of this instrument can be found elsewhere (Spitzer, Endicott, & Cohen, 1964; Spitzer, Fleiss, Burdock, & Hardesty, 1964; Spitzer, Fleiss, Endicott, & Cohen, 1967).

The advantages of the MSS over other commonly used assessment procedures include the incorporation of a standardized interview schedule to reduce in-

consistency and oversight due to variability in interviewing technique and coverage of psychopathology, awareness of what questions are asked to provide a framework within which the patient's responses can be understood by others not present at the interview, and the use of a score sheet that serves simultaneously as a permanent clinical record and as a form for automated data processing. The use of the same interview schedule for all patients also has the research advantage that differences observed among patients tend to reflect actual differences rather than artifacts caused by differences in areas of psychopathology explored or interviewing techniques used.

Procedure

Subjects were tested approximately 48 to 72 hours after admission as part of the routine screening procedure. A counterbalanced design was used to offset evidence of decreased pathology under repeated MMPI administrations (Kincannon, 1968; Newton, 1971; Windle, 1954). The subjects were assigned alternatively to one of two groups: one group initially received the MMPI, and the other received either the FAM or the MMPI-168. As soon as possible after completion of this initial testing, the procedure was reversed so that each subject received both the standard and one abbreviated MMPI form. Thus, independently administered abbreviated forms were obtained, since the extracted method possessed inherent flaws. Most notably, a perfect reliability of scores was assumed, since the same items scored were used on both the abbreviated and standard forms. Furthermore, Newmark et al. (1973) presented evidence that the extracted form has limited benefits.

The FAM was scored and converted into standard-scale raw scores using Fashingbauer's (1973) tables of conversion for each sex. Conversion of prorated raw scores to *T* scores was then accomplished following the standard procedures. The MMPI-168 was scored and converted into standard-scale *K*-corrected raw scores using regression equations for each scale as derived by Overall and Gomez-Mont (1974). *K*-corrected *T* scores were used in this study in order to be consistent with the literature. The standard-form MMPI answer sheets were hand scored in the traditional manner. Invalidity was defined as either $F \geq 90$, L or $K \geq 70$ or $? \geq 60$.

Four raters, consisting of two clinical psychologists and two psychiatric residents, were presented detailed operationally defined explanations of the symptoms on the BPRS as defined by Overall and Gorham (1962) and Porkorny and Overall (1970) in an attempt to reduce idiosyncratic biases. Whereas definitions can never be rigorous or complete except in mathematics, they nevertheless serve to demarcate a concept even though its boundaries remain somewhat blurred (Zubin, 1967). All raters had at least 4 years of diagnostic experience. Such a procedure was necessary because it has been demonstrated (Kreitman, 1961) that variables relating to

nomenclature and degree of experience were the greatest impediments to reliability in psychiatric diagnoses. Clearly definable terms and equivalent diagnostic experience are essential.

Initially, in an attempt to maximize interrater reliability, filmed MSS interviews with three psychiatric inpatients, with varied degrees and types of psychopathology, were presented to the four raters. Following each filmed presentation a detailed discussion of the rating differences occurred. That such a procedure has merit in increasing rater reliability has been confirmed by Raskin, Schulterbrandt, and Reatig (1966).

Each patient was then interviewed using the MSS by one of the raters while another rater observed this initial diagnostic interview behind a one-way mirror. Neither rater was familiar with the patient's history or observed ward behavior. Reliability is definitely enhanced if two raters observe the same interview rather than each rater observing a separate interview (Wittenborn, 1972). Thus, each patient was rated on the BPRS by two raters so that interrater reliability was assessed. Immediately following the interview, each rater recorded his observations independent of his colleague's rating.

When two raters observe the same patient, two distinct strategies exist for combining data (Overall, 1968). One approach has been to have both observers discuss each rating at the conclusion of the interview and to arrive at a consensus rating. The other more frequently used model has been to have raters complete ratings independently and then to average the ratings on the theory that random errors tend to cancel. This latter model has been

shown to be an advantageous procedure by Overall, Hollister, and Dalal (1967). Consequently, the mean ratings for each BPRS symptom were used.

Since several MMPI scales contain more than one interpretative factor, these scales were correlated with the means of more than one BPRS symptom rating using multiple correlation coefficients (Bruning & Kintz, 1968). Only for Scale *L*, which was correlated with only one BPRS symptom, was a product-moment correlation used. The MMPI scales and the corresponding BPRS scales with which they were correlated were as follows: *L*-Denial; *F*-Unusual Thought Content, Conceptual Disorganization, Intellectual Subnormality; *K*-Denial, Guardedness; *Hs*-Conversion Somatization, Somatic Concern; *D*-Depressive Mood, Introjection of Blame; *Hy*-Conversion Somatization, Dramatization, Manipulativeness, Denial, Emotionally Labile; *Pd*-Antisocial Trends, Impulsiveness, Projection of Blame, Hostility, Manipulativeness; *Mf*-Sexual Crossed Identification, Passive Dependency; *Pa*-Suspiciousness, Projection of Blame, Hostility; *Pt*-Anxiety, Tension, Guilt Feelings, Phobias, Compulsive Acts, Obsessive Thoughts; *Sc*-Conceptual Disorganization, Unusual Thought Content, Sexual Inadequacy, Emotional Withdrawal, Blunted Affect, Feelings of Unreality; and *Ma*-Excitement, Elevated Mood, Grandiosity, Tension. The BPRS ratings selected for inclusion were based on item analyses content and interpretive data of the MMPI scales as presented by Dahlstrom, Welsh, and Dahlstrom (1972), Fowler (1966), and Lachar (1974).

Table 1
Mean T Scores, Standard Deviations, Correlations, and t Values for the Standard MMPI and the Faschingbauer Abbreviated MMPI (FAM)

Scale	Males (n = 60)						Females (n = 85)					
	MMPI		FAM		r	t	MMPI		FAM		r	t
	M	SD	M	SD			M	SD	M	SD		
<i>L</i>	49.1	8.4	49.7	8.5	.82	2.87*	49.8	7.4	50.0	7.7	.85	1.42
<i>F</i>	66.7	12.8	67.2	12.8	.88	1.95	67.7	12.8	68.3	12.9	.95	3.74**
<i>K</i>	51.4	10.8	51.5	10.9	.93	.89	50.1	8.3	50.4	8.4	.86	1.93
<i>Hs</i>	63.2	14.0	63.1	14.1	.93	.37	61.6	13.3	62.1	13.2	.95	3.17*
<i>D</i>	75.5	15.9	75.5	15.7	.92	.13	72.1	13.9	72.4	14.1	.95	1.85
<i>Hy</i>	65.3	11.3	65.2	11.5	.91	.41	66.9	11.6	67.2	11.7	.94	1.97
<i>Pd</i>	72.6	12.4	72.3	12.1	.84	1.19	72.4	13.4	73.8	11.5	.83	1.62
<i>Mf</i>	63.7	11.0	63.4	11.1	.87	1.70	46.9	12.2	47.1	12.3	.94	1.23
<i>Pa</i>	67.6	11.7	67.5	11.5	.89	.34	69.1	11.8	69.7	12.2	.90	3.29*
<i>Pt</i>	72.2	13.8	72.7	13.6	.90	1.87	69.3	11.2	69.8	11.4	.93	3.42**
<i>Sc</i>	75.5	15.5	76.0	15.6	.89	1.92	73.6	13.8	73.5	14.9	.90	.41
<i>Ma</i>	61.5	11.8	61.7	11.9	.90	1.09	62.5	12.7	62.7	12.9	.95	1.35

Note. MMPI = Minnesota Multiphasic Personality Inventory.
* $p \leq .01$.
** $p \leq .001$.

Results

The *T*-score means, standard deviations, Pearson product-moment correlations, and *t* values of the comparable validity, clinical, and *Mf* scales of both the MMPI and the FAM are presented as a function of sex in Table 1. Note that in most cases for both males and females the standard deviation was larger for the FAM than for the comparable standard MMPI scales. Statistical analysis showed this to be a significantly reliable trend ($p \leq .05$).

Paired *t* tests yielded significant mean differences on *L*, $t(59) = 2.87$, $p \leq .01$, for males, and on *F*, $t(84) = 3.74$, $p \leq .001$; *Hs*, $t(84) = 3.17$, $p \leq .01$; *Pa*, $t(84) = 3.29$, $p \leq .01$, and *Pt*, $t(84) = 3.42$, $p \leq .001$, for females. In all cases the FAM significantly overestimated the MMPI scale scores. This finding is consistent with previously published results (e.g., Newmark et al., 1973), demonstrating that the FAM seemed to overestimate slightly MMPI scale means.

For males, scale correlations ranged from .82 for *L* to .93 for *K* and *Hs* ($M = .88$, $Mdn = .89$). For females, correlations ranged from .83 on *Pd* to .95 on *F*, *Hs*, *D*, and *Ma* ($M =$

.91, $Mdn = .93$). All correlations were significantly different from zero ($p \leq .001$).

The *T*-score means, standard deviations, and Pearson product-moment correlations of the comparable validity, clinical, and *Mf* scales of both the MMPI and the MMPI-168 are presented as a function of sex in Table 2. Note that in approximately half of the cases for each sex, the standard deviation was larger for the MMPI-168 than for the comparable standard MMPI scale.

Paired *t* tests yielded no significant mean differences for male subjects. For females, however, significant mean differences occurred on *F*, $t(84) = 4.29$, $p \leq .001$; *Pd*, $t(84) = 3.09$, $p \leq .01$; *Pa*, $t(84) = 2.96$, $p \leq .01$; and *Sc*, $t(84) = 4.59$, $p \leq .001$. In all cases the MMPI-168 significantly overestimated the MMPI scale scores. The results for females are rather surprising in view of several previously published reports (Newmark, Newmark, & Cook, 1975; Newmark & Raft, 1976) that found no significant mean differences for a sample of both male and female psychiatric inpatients and medical patients.

For males, scale correlations ranged from .92 on *L* to .98 on *Hs* and *Sc* ($M = .96$, $Mdn = .96$). For females, scale correlations ranged

Table 2
Mean *T* Scores, Standard Deviations, Pearson Correlations, and *t* Values for the Standard MMPI and the MMPI-168

Scale	Males ($n = 60$)						Females ($n = 85$)					
	MMPI		MMPI-168		<i>r</i>	<i>t</i>	MMPI		MMPI-168		<i>r</i>	<i>t</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
<i>L</i>	49.8	8.5	50.0	8.4	.92	1.26	51.1	8.9	51.4	8.8	.90	1.89
<i>F</i>	68.3	15.0	68.4	15.1	.96	.80	68.1	12.5	68.8	12.4	.93	4.29**
<i>K</i>	51.7	10.2	51.5	10.1	.94	1.41	50.6	8.9	50.7	9.0	.89	.58
<i>Hs</i>	65.7	16.0	65.5	16.0	.98	1.35	64.8	11.8	65.0	11.9	.94	1.78
<i>D</i>	77.2	17.4	77.1	17.3	.97	.77	73.9	15.1	74.1	15.0	.96	1.51
<i>Hy</i>	69.2	12.5	68.9	12.3	.96	1.59	70.1	11.0	70.2	11.1	.94	1.07
<i>Pd</i>	74.8	12.6	74.7	12.5	.95	.62	74.3	12.2	74.8	12.0	.93	3.09*
<i>Mf</i>	66.7	10.5	66.5	10.4	.94	.88	47.5	9.4	47.7	9.2	.90	1.21
<i>Pa</i>	70.0	13.5	70.1	13.6	.96	.30	70.3	11.6	70.8	11.7	.92	2.96*
<i>Pt</i>	75.1	16.8	74.9	16.9	.97	1.26	72.0	12.1	72.1	12.3	.94	.58
<i>Sc</i>	78.7	20.3	78.6	20.4	.98	.77	76.3	16.1	77.2	16.3	.95	4.59**
<i>Ma</i>	66.7	13.3	66.8	13.4	.97	.35	62.9	13.8	63.1	14.2	.95	1.22

Note. MMPI = Minnesota Multiphasic Personality Inventory.

* $p \leq .01$.

** $p \leq .001$.

Table 3

Correlations Between the BPRS Ratings and the Corresponding MMPI and Short-Form Scale Scores

Scale	Males		Females		Males		Females	
	MMPI	FAM	MMPI	FAM	MMPI	MMPI-168	MMPI	MMPI-168
<i>L</i>	.63	.63	.57	.54	.64	.63	.61	.64
<i>F</i>	.74	.72	.74	.73	.70	.71	.69	.68
<i>K</i>	.80	.81	.78	.79	.65	.67	.81	.79
<i>Hs</i>	.82	.82	.83	.86	.79	.80	.84	.86
<i>D</i>	.83	.85	.84	.81	.81	.81	.83	.83
<i>Hy</i>	.86	.85	.83	.83	.82	.84	.83	.86
<i>Pd</i>	.83	.80	.65	.79*	.77	.79	.82	.79
<i>Mf</i>	.85	.82	.80	.78	.83	.84	.75	.77
<i>Pa</i>	.77	.79	.82	.80	.83	.84	.83	.81
<i>Pt</i>	.80	.79	.76	.76	.81	.80	.81	.81
<i>Sc</i>	.61	.64	.72	.69	.65	.67	.64	.66
<i>Ma</i>	.85	.85	.81	.82	.84	.83	.86	.84

Note. BPRS = Brief Psychiatric Rating Scale; MMPI = Minnesota Multiphasic Personality Inventory; FAM = Fashingbauer's Abbreviated MMPI. $n = 60$ for males and 85 for females.

* $p \leq .01$.

from .89 on *K* to .96 on *D* ($M = .93$, $Mdn = .93$). All correlations were significantly different from zero ($p \leq .001$).

Table 3 presents the multiple correlation coefficients between the BPRS ratings and the corresponding MMPI and abbreviated-form scales. All of the correlations were significantly different from zero ($p \leq .001$). For males, correlations ranged from .61 for *Sc* to .86 for *Hy* ($M = .78$, $Mdn = .81$) on the standard form and from .63 for *L* to .85 for *D*, *Hy*, and *Ma* ($M = .80$, $Mdn = .82$) on the FAM. For females, correlations ranged from .57 for *L* to .84 for *D* ($M = .78$, $Mdn = .79$) on the standard form and from .54 for *L* to .86 for *Hs* ($M = .79$, $Mdn = .79$) on the FAM. With regard to the MMPI and the MMPI-168, for males, correlations ranged from .64 for *L* to .84 for *Ma* ($M = .78$, $Mdn = .81$) on the standard form and from .63 for *L* to .84 for *Hy*, *Mf*, and *Pa* ($M = .79$, $Mdn = .79$) on the MMPI-168. For females, correlations ranged from .61 for *L* to .86 for *Ma* ($M = .79$, $Mdn = .82$) on the standard form and from .64 for *L* to .86 for both *Hs* and *Hy* ($M = .79$, $Mdn = .81$) for the MMPI-168. Only on *Pd* for females did a significant difference occur between the MMPI and the abbreviated-form correlations, $t(84) = 2.72$, $p \leq .01$, with the FAM being

significantly higher. However, this significant difference could occur by chance, since 48 comparisons were made.

Paired *t* tests between the multiple correlation coefficients from the BPRS ratings and each abbreviated form yielded significantly higher correlations for females on *L*, $t(84) = 3.05$, $p \leq .01$; *F*, $t(84) = 2.32$, $p \leq .05$; and *Pt*, $t(84) = 2.07$, $p \leq .05$. For *L* and *Pt*, the MMPI-168 correlation coefficients were higher, whereas for *F* the FAM correlation coefficients were higher. For males, significantly higher correlations occurred for *K*, $t(59) = 2.66$, $p \leq .01$, on the FAM and *Pa*, $t(59) = 2.19$, $p \leq .05$, on the MMPI-168.

Product-moment reliability coefficients across the four raters ranged from .83 to .88 ($M = .86$). Mean interrater reliabilities between the BPRS scores and the corresponding MMPI scale scores were as follows: *L* = .87, *F* = .86, *K* = .84, *Hs* = .88, *D* = .88, *Hy* = .87, *Pd* = .84, *Mf* = .87, *Pa* = .83, *Sc* = .85, and *Ma* = .88. Each rater rated the same number of patients.

Discussion

Although both abbreviated forms significantly overestimated several of the comparable standard MMPI scale means, the actual

mean differences that occurred were extremely small, ranging from .3 to .9, and probably were not of any practical significance. Hill (1976) presented compelling evidence, suggesting that the concepts of statistical validity and clinical utility were by no means identical and that it does not necessarily follow that one is a prerequisite for the other. Furthermore, related *t* tests are concerned with the magnitude of the correlation so that when large correlations, as were obtained in this study, are found, the differences between the means can be extremely small and still be statistically significant.

The majority of abbreviated-form investigations have demonstrated a tendency for these forms to underestimate MMPI *T*-score standard deviations, leading Dean (1972) to suggest that this is somehow inherent in shorter tests. However, Faschingbauer (1976) concluded that his use of a substitution model rather than a regression model in developing conversion values for the FAM would overcome this persistent tendency. Equivocal support for this hypothesis was found in the present investigation, as the FAM had generally larger standard deviations than did the standard-form MMPI, whereas the MMPI-168, which was developed using a regression model, had higher standard deviations for approximately half of the scales for each sex.

The correlations obtained between the standard MMPI and the BPRS ratings were markedly greater than results obtained from other investigations (Endicott & Jortner, 1966, 1967; Endicott, Jortner, & Abramoff, 1969; Harris, Wittner, Koppell, & Hilf, 1970), which correlated other objective measures of psychopathology with various standard MMPI scales. The higher correlations may have occurred in the present investigation due to the use of multiple correlation coefficients, which combine several objective measures of each MMPI scale rather than using just one measure. For example, instead of correlating *Pt* with objective measures of anxiety only, a multiple correlation coefficient was obtained from objective measures of anxiety, tension, guilt, phobias, compulsive acts, and obsessive thoughts.

The correlations obtained between the ab-

breivated forms and the BPRS ratings were significantly high and comparable to the MMPI-BPRS rating correlations. Scale *L* correlations were consistently lower in all cases, possibly because only one objective measure of psychopathology was correlated with it. Only on *Pd* for females did a significant difference occur between the MMPI and the abbreviated-form correlations, with the FAM being significantly higher. However, this difference quite likely could have occurred by chance. Thus, the empirical validity of both abbreviated forms seems comparable to the standard-form MMPI when compared with direct measures of psychopathology.

The FAM-BPRS rating correlations and the MMPI-168-BPRS rating correlations were significantly different in 5 of the 24 comparisons made. However, neither abbreviated form demonstrated any superiority.

The interrater reliabilities obtained in this investigation were comparable to those obtained in the original developmental investigation of the BPRS by Overall and Gorham (1962). However, the present results were somewhat higher than those obtained in the majority of recent BPRS investigations (e.g., Anderson, Kuehnle, & Catanzano, 1976) possibly due to the attempt to maximize interrater reliability through the use of practice procedures as advocated by Raskin et al. (1966).

As an additional analysis, the mean BPRS symptoms rating scores were obtained for each comparable MMPI scale. Both abbreviated- and standard-form MMPIs in each sample as a function of sex were then subgrouped using the mean BPRS ratings as a moderator variable. Means between 0 and 2.5 were included in the low-ratings group, and means between 3.5 and 6.0 were included in the high-ratings group. Means from 2.6 to 3.4 were excluded to provide more well-defined groups. The mean MMPI scale scores for the low-ratings group and high-ratings group were not significantly different from the mean abbreviated-form scales for the low-ratings and high-ratings groups, respectively, within both samples regardless of sex.

Endinger, Kendall, Hooke, and Bogan (1976) emphasized that in the absence of in-

dependent validation, none of the abbreviated forms developed can be recommended for clinical use. However, there is evidence from the present investigation that both the FAM and MMPI-168 were comparable to the standard-form MMPI when compared with direct measures of psychopathology. Since both abbreviated forms are slightly less expensive to administer and result in more completed profiles with no loss in empirical validity, they appear to be an adequate substitute when administration of the standard-form MMPI is not feasible.

Reference Note

1. Spera, J., & Robertson, M. A 104-item MMPI: The Maxi-Mult. Paper presented at the meeting of the American Psychological Association, New Orleans, September 1974.

References

- Anderson, W. H., Kuehne, J. C., & Catanzano, D. N. Rapid treatment of acute psychosis. *American Journal of Psychiatry*, 1976, 133, 1076-1078.
- Auerbach, A. H., & Ewing, J. H. Some limitations of psychiatric rating scales: The clinician's viewpoint. *Comprehensive Psychiatry*, 1964, 5, 93-99.
- Bruning, J. L., & Kintz, B. L. *Computational handbook of statistics*. Glenview, Ill.: Scott, Foresman, 1968.
- Cleveland, S. E. Reflections on the rise and fall of psychodiagnosis. *Professional Psychology*, 1976, 7, 309-318.
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook* (Vol. 1). *Clinical interpretation*. Minneapolis: University of Minnesota Press, 1972.
- Dean, E. F. A lengthened Mini: The Midi-Mult. *Journal of Clinical Psychology*, 1972, 28, 68-71.
- Endicott, N. A., & Jortner, S. Objective measures of depression. *Archives of General Psychiatry*, 1966, 15, 249-255.
- Endicott, N. A., & Jortner, S. Correlates of somatic concern derived from psychological tests. *Journal of Nervous and Mental Disease*, 1967, 144, 133-138.
- Endicott, N. A., Jortner, S., & Abramoff, E. Objective measures of suspiciousness. *Journal of Abnormal Psychology*, 1969, 74, 26-32.
- Endinger, J. D., Kendall, P. C., Hooke, J. F., & Bogan, J. B. The predictive efficacy of three MMPI short forms. *Journal of Personality Assessment*, 1976, 40, 259-265.
- Faschingbauer, T. R. A short written form of the group MMPI. (Doctoral dissertation, University of North Carolina, 1972). *Dissertation Abstracts International*, 1973, 34, 409B. (University Microfilms No. 73-16, 469)
- Faschingbauer, T. R. Substitution and regression models, base rates and the clinical validity of the Mini-Mult. *Journal of Clinical Psychology*, 1976, 32, 70-74.
- Fowler, R. D. *The MMPI notebook: A guide to the clinical use of the automated MMPI*. Nutley, N.J.: Roche Psychiatric Service Institute, 1966.
- Goldberg, L. R. Objective diagnostic tests and measures. *Annual Review of Psychology*, 1974, 25, 343-366.
- Harris, R. J., Wittner, W., Koppell, B., & Hilf, F. D. MMPI scales vs. interviewer ratings of paranoia. *Psychological Reports*, 1970, 27, 447-450.
- Hill, A. D. Validity and clinical utility of the Grid Test of Schizophrenic Thought Disorder. *British Journal of Psychiatry*, 1976, 126, 251-254.
- Hugo, J. A. Abbreviation of the MMPI through multiple regression. (Doctoral dissertation, University of Alabama, 1971). *Dissertation Abstracts International*, 1971, 32, 1213B. (University Microfilms No. 71-20, 115)
- Kincannon, J. C. Prediction of the standard MMPI scale scores from 71 items: The Mini-Mult. *Journal of Consulting and Clinical Psychology*, 1968, 32, 319-325.
- Kreitman, N. The reliability of psychiatric diagnosis. *Journal of Mental Science*, 1961, 107, 876-886.
- Lachar, D. *The MMPI: Clinical assessment and automated interpretation*. Los Angeles: Western Psychological Services, 1974.
- Lorr, M. Ratings scales and checklists for the evaluation of psychopathology. *Psychological Bulletin*, 1954, 51, 119-127.
- Newmark, C. S., Boas, B., & Messervy, T. An abbreviated MMPI for use with college students. *Psychological Reports*, 1974, 34, 631-634.
- Newmark, C. S., Conger, A. J., & Faschingbauer, T. R. The interpretive validity and effective test length functioning of an abbreviated MMPI relative to the standard MMPI. *Journal of Clinical Psychology*, 1976, 32, 27-32.
- Newmark, C. S., Cook, L., Clarke, M., & Faschingbauer, T. Application of Faschingbauer's abbreviated MMPI to psychiatric inpatients. *Journal of Consulting and Clinical Psychology*, 1973, 41, 416-421.
- Newmark, C. S., Falk, R., & Finch, A. J. Interpretive accuracy of abbreviated MMPIs. *Journal of Personality Assessment*, 1976, 40, 266-268.
- Newmark, C. S., Galen, R., & Gold, K. Efficacy of an abbreviated MMPI as a function of type of administration. *Journal of Clinical Psychology*, 1975, 31, 639-642.
- Newmark, C. S., & Glenn, L. Sensitivity of the Faschingbauer Abbreviated MMPI to hospitalized adolescents. *Journal of Abnormal Child Psychology*, 1974, 2, 299-306.
- Newmark, C. S., Newmark, L. G., & Cook, L. The MMPI-168 with psychiatric patients. *Journal of Clinical Psychology*, 1975, 31, 61-64.
- Newmark, C. S., Newmark, L., & Faschingbauer, T. R. Utility of three abbreviated MMPIs with psy-

- chiatric outpatients. *Journal of Nervous and Mental Disease*, 1974, 159, 438-443.
- Newmark, C. S., Owen, M., Newmark, L., Cook, L., & Faschingbauer, T. R. Comparison of three abbreviated MMPIs for psychiatric patients and normals. *Journal of Personality Assessment*, 1975, 39, 261-270.
- Newmark, C. S., & Raft, D. Using an abbreviated MMPI as a screening device for medical patients. *Psychosomatics*, 1976, 17, 45-48.
- Newton, J. R. A comparison of studies of the Mini-Mult. *Journal of Clinical Psychology*, 1971, 27, 489-490.
- Norton, W. A. A review of psychiatric rating scales. *Canadian Psychiatric Association Journal*, 1967, 12, 563-574.
- Overall, J. E. Standard psychiatric symptom description: The Factor Construct Rating Scale (FCRS). *Triangle*, 1968, 8, 178-184.
- Overall, J. E. The Brief Psychiatric Rating Scale in psychopharmacology research. *Modern Problems of Pharmacopsychiatry*, 1974, 7, 67-68.
- Overall, J. E., & Gomez-Mont, F. The MMPI-168 for psychiatric screening. *Educational and Psychological Measurement*, 1974, 34, 315-319.
- Overall, J. E., & Gorham, D. R. The Brief Psychiatric Rating Scale. *Psychological Reports*, 1962, 10, 799-812.
- Overall, J. E., Higgins, W., & De Schweinitz, A. Comparison of differential diagnostic discrimination for abbreviated and standard MMPI. *Journal of Clinical Psychology*, 1976, 32, 237-245.
- Overall, J. E., Hollister, L. E., & Dalal, S. N. Psychiatric drug research: Sample size requirements for one versus two rates. *Archives of General Psychiatry*, 1967, 16, 152-161.
- Overall, J. E., Hollister, L. E., Johnson, M., & Pennington, V. Nosology of depression and differential response to drugs. *Journal of the American Medical Association*, 1966, 195, 946-948.
- Overall, J. E., Hollister, L. E., & Pichot, P. Major psychiatric disorders. *Archives of General Psychiatry*, 1967, 16, 146-151.
- Overall, J. E., & Klett, C. J. *Applied multivariate analysis*. New York: McGraw-Hill, 1972.
- Pokorny, A. E., & Overall, J. E. Relationships of psychopathology to age, sex, ethnicity, education and marital status in state hospital patients. *Journal of Psychiatric Research*, 1970, 7, 143-152.
- Rand, S. A comprehensive validation of the expanded Mini-Mult. Unpublished master's thesis, University of Miami, 1976.
- Raskin, A., Schulerbrandt, J. G., & Reatig, N. Rater and patient characteristics associated with rater differences in psychiatric scale ratings. *Journal of Clinical Psychology*, 1966, 22, 417-423.
- Schreier, A. J., Reznikoff, M., & Glueck, B. C. Controlled clinical evaluation of acute schizophrenia: A comparative study of rating instruments and judges. *Comprehensive Psychiatry*, 1962, 3, 343-353.
- Spitzer, R. L., Burdock, E. I., & Hardesty, A. S. *Mental Status Schedule*. New York: Biometrics Research, New York State Department of Mental Hygiene, 1964.
- Spitzer, R. L., Endicott, J., & Cohen, G. M. *Manual of instruction for the Mental Status Schedule*. New York: Biometrics Research, New York State Department of Mental Hygiene, 1964.
- Spitzer, R. L., Fleiss, J. L., Burdock, E. I., & Hardesty, A. S. The Mental Status Schedule: Rational, reliability and validity. *Comprehensive Psychiatry*, 1964, 5, 384-393.
- Spitzer, R. L., Fleiss, J. L., Endicott, J., & Cohen, J. Mental Status Schedule: Properties of factor-analytically derived scales. *Archives of General Psychiatry*, 1967, 16, 479-492.
- Spitzer, R. L., Fleiss, J. L., Kernohan, W., Lee, J. C., & Baldwin, I. T. The Mental Status Schedule: Comparing Kentucky and New York schizophrenics. *Archives of General Psychiatry*, 1965, 12, 448-455.
- Steer, R. A. Relationship between the Brief Psychiatric Rating Scale and the Multiple Affect Adjective Checklist for schizophrenic women. *Psychological Reports*, 1974, 35, 79-82.
- Windle, C. Further studies of test-retest effect on personality questionnaires. *Educational and Psychological Measurement*, 1954, 15, 246-253.
- Wittenborn, J. R. Reliability, validity, and objectivity of symptom-rating scales. *Journal of Nervous and Mental Disease*, 1972, 154, 79-87.
- Zubin, J. Classification of the behavior disorders. *Annual Review of Psychology*, 1967, 18, 373-403.

Received October 3, 1976 ■

Treatment of Female Sexual Dysfunction Through Symbolic Modeling

Georgia H. Nemetz, Kenneth D. Craig, and Gunther Reith
University of British Columbia, Vancouver, Canada

An evaluation of treatment programs for women suffering from debilitating sexual anxiety is described. Attitudinal and behavioral indices of sexual adjustment and sexual anxiety were obtained from 22 women to assess effects of individual and group graduated symbolic modeling through videotapes, with concurrent behavioral tasks as treatment procedures. All women serving as clients had reported severe anxiety that precluded sexual enjoyment or activity. Sixteen clients were randomly assigned to two groups receiving either individual or group treatment. Treatment consisted of relaxation training followed by viewing 45 videotaped vignettes depicting graduated sexual behaviors. Five sessions were held twice weekly. The other 6 women were subjected to identical measurement without the benefit of treatment. Decreases in anxiety and increases in behavioral and attitudinal measures were evidenced; however, a trend toward greater improvement was observed for those receiving group treatment. Improvement remained stable through a 1-year follow-up. Control clients showed no improvement and trends toward deterioration. The treatment regimen appeared to produce positive changes in attitudinal, behavioral, and anxiety indices simultaneously.

Sexual anxiety continues to be a major concern for therapists concerned with sexual dysfunction and anomalies. Although a diversity of therapeutic techniques have been devised to accommodate the numerous theoretical positions (Kaplan, 1974; Masters & Johnson, 1970), there is general agreement that anxiety toward sexual functioning is a basic prohibitive factor in the enactment and enjoyment of sexual activity. As anxiety, whether defined or measured affectively, behaviorally, or physiologically, is a highly learnable pattern of response, aversive sexual incidents would be sufficient to generate female sexual dysfunction. These "incidents" can range from rape and parental molestation to anxiety-inducing instruction from others, failure to achieve "expected orgasm," and pre-occupying concern with physical attributes. For instance, Masters and Johnson (1970) conceptualized performance anxiety and spec-

tating as the two major inhibiting factors to the natural flow of stimuli leading to the female orgasmic response. Because anxiety is readily attached to previously noneliciting stimuli on the basis of simple contiguity, the full range of sexual behavior may become affected. For example, anxiety toward intercourse may generalize sufficiently to disrupt or preclude other effective sexual habits.

A number of therapists have effectively used the desensitization paradigm (Wolpe, 1958) to reduce sexual anxiety and achieve a concurrent increase in pleasurable sexual behavior (Brady, 1966; Kraft & Al-Issa, 1967; Lazarus, 1963). Lazarus (1963) treated 16 inorgasmic women described as recalcitrant and persistent cases. Subjects were instructed in progressive relaxation, hierarchies with respect to sexual functioning were constructed, and anxiety-eliciting items were presented verbally in a hierarchical fashion. Lazarus reported that 9 patients became "sexually adjusted" after a mean of 28.7 sessions over an average period of 6 months. Control group data were not provided. Kraft and Al-Issa (1967) used desensitization for a

Requests for reprints should be sent to Georgia H. Nemetz, Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5.

female patient whose dysfunction was attributed partially to social anxiety in heterosexual situations. The authors reported complete recovery after 84 1½-hour systematic desensitization sessions. Variations in traditional desensitization procedures have also proven effective. Clopton and Risbrough (1973) reported success in eliminating sexual anxiety and in inducing approach behavior by having patients present sexually aversive stimuli to themselves at or before masturbation-induced orgasm. Brady (1966) reported success for 4 out of 5 clients suffering from "chronic severe frigidity" using desensitization aided by the use of subanesthetic doses of methohexital sodium as a means of producing profound muscular relaxation.

The therapeutic procedures used in the present investigation were based on modeling strategies because of their effectiveness in reducing maladaptive fears and avoidance behavior and in concomitantly inducing desirable behavior. Elements of vicarious extinction (Bandura, Grusec, & Menlove, 1967; Craig, Best, & Ward, 1975), graduated symbolic modeling (Bandura, Blanchard, & Ritter, 1969), and videotaped desensitization (Wincze & Caird, 1976; Woody & Schauble, 1969) were utilized. Potential effects of modeling therapies, previously identified (Bandura, 1969), include the acquisition of previously nonexistent patterns of behavior, weakening response inhibitions, and facilitating the occurrence of preexisting responses in the behavior repertoire. In the instance of sexual anxiety, modeling therapy would appear to be particularly promising because of its effectiveness in extinguishing anxiety and avoidance tendencies and in inducing behavioral and attitudinal change. Bandura et al. (1969) found graduated symbolic modeling to be more effective than standard desensitization in neutralizing the anxiety-arousing properties of snakes and in inducing the greatest behavioral and attitudinal change.

The present study used symbolic modeling through the medium of a videotaped, graduated hierarchy of sexual scenes. Concurrent behavioral tasks also were utilized in the form of homework assignments to match the

scenes portrayed in the videotapes. It was expected that reducing sexual anxiety by vicarious extinction would enable the client to successfully complete the behavioral tasks, thus exposing them to a potentially positive sexual experience. The treatment procedures were based on those used by Wincze and Caird (1976), who found that videotaped desensitization was more effective than standard desensitization in reducing heterosexual anxiety in those suffering from primary sexual dysfunction. The present study extended these treatment procedures to a group therapy program, used additional measurement procedures to examine other components of the behavior change process, and provided a 1-year follow-up of treatment effectiveness.

In addition to the potential that symbolic modeling techniques have for effective intervention, they introduce economies by reducing requirements for therapist involvement—thus making therapy more readily available. In the extreme, the technique can be exclusively self-administered. The procedure also lends itself to application to several clients at the same time. We could find no accounts of videotape treatment used in a group format in the literature. The present study contrasted individual treatment, group treatment, and a waiting list control group.

A multidimensional model for evaluating treatment effectiveness was adopted. Changes in subjective affect were evaluated through self-report scales. Behavioral indices of changes in sexual activities were provided by clients who self-monitored during baseline, treatment, and follow-up phases and by the clients' male partners, who also provided data during these phases of the treatment program. Repetition of the measures, especially the behavioral indices, was designed to assess how rapidly treatment effects would manifest themselves. Because of the potential for producing a reorganization of attitudes toward personal and general sexual behavior, scales evaluating this domain also were included.

Method

Subjects

The subjects were 22 inorgasmic female clients, 15 of whom would satisfy Masters and Johnson's

(1970) criteria for secondary orgasmic dysfunction and 7 of whom would be classified as primary inorganic. All had been referred by physicians to the Sexual Dysfunction Clinic at the Health Sciences Center Hospital at the University of British Columbia. Other selection criteria included presence of anxiety or discomfort toward sexual behavior as measured both by interview data and scores on the Sexual Anxiety Card Sort (Barlow, Leitenberg, & Agras, Note 1); absence of organic causal factors, as established through gynecological examination; absence of acute marital problems justifying inference of an unstable relationship with the sexual partner; and availability and cooperation of the husband or a steady male partner. Of 29 women initially interviewed, 7 failed to score above a cutoff point of 13 out of 100 on the Sexual Anxiety Card Sort. The subjects ranged in age from 21 to 39 years with a median of 26.7 years. Fifteen of the subjects were married, 4 were single, and 2 were separated. The first 16 subjects obtained by successive admissions to the clinic were randomly assigned to one of the two treatment conditions. The last 6 successive admissions were assigned to the control condition. These control subjects were advised truthfully that treatment was not available for approximately 6 weeks but that it would be provided on termination of other clients. All 6 volunteered to participate in pretreatment investigations and provided data on the measures matching those used by experimental subjects. At the end of this period of time, all were treated in this clinic.

Measures

Measures of anxiety included (a) the sexual anxiety Card Sort (Barlow et al., Note 1), which consists of 25 separate cards, describing various sexual behaviors, ranging from mere physical proximity to intercourse and orgasm, that are sorted by the client according to the levels of anxiety induced if she were to engage in the depicted behaviors. The cards are sorted into five categories with scores assigned to them ranging between 0 and 4. Initial admission scores for clients ranged between 13 and 70 with a median of 32. This measure was administered at the initial interview, prior to each of the five treatment sessions, at the 3-week follow-up, and at the 1-year follow-up. (b) On the Bentler Heterosexual Behavior Hierarchy (Bentler, 1968), clients rate 25 statements describing sexual behavior on a scale of 0 to 4 with the extremes labeled *no anxiety* and *very much anxiety*. As this measure contains novel sexual items not depicted in the modeling hierarchy, it was viewed as an indicant of treatment generalization. The Bentler scale was administered at the initial interview, at the 3-week follow-up, and at the 1-year follow-up.

The behavioral measures were based on a sexual behavior index, constructed for this study to measure the incidence of specific sexual behaviors. This instrument requires a recording of the frequency of

nine explicit categories of sexual behavior (see Table 1). Both partners independently record the frequency of their sexual behaviors corresponding to these categories for complete weeks, including 1 week prior to treatment, the intervals between each treatment, during the 3-week follow-up period, and during the week prior to the 1-year follow-up.

Attitudinal measures examining both global and specific sexual attitudes were (a) Rotter's Sex Attitude Scale (Rotter, Note 2), which contains 100 items requiring ratings of attitudes on sexual issues concerning sex roles, sex education, and marital relationships. The measure was administered at the initial interview, prior to the fifth treatment session, and at the 3-week follow-up. (b) The Sexual Semantic Differential assesses attitudes toward specific sexual concepts. Nine of the bipolar scales were based on the Marks and Sartorius (1968) measure. Ratings were scored for two basic dimensions, evaluation and anxiety. Clients rated seven concepts including both specific sexual activities and basic sexual descriptors (see Table 1). The semantic differential was administered at the initial interview, each of the five treatment sessions, the 3-week follow-up, and the 1-year follow-up.

Procedure

At an initial interview, after acceptance into the treatment program, both partners were advised of procedures and regulations. The male's supportive potential was established, and he was instructed to refrain from initiating sexual activity for the duration of treatment. The different measures were described, and questionnaires were completed. The rationale for relaxation training was explained to the client, and she received a tape-recorded or card-form summary of progressive relaxation procedures. She was instructed to practice the exercises twice a day during sessions of approximately 15 minutes.

Five sessions twice a week with 3-day intersession intervals were planned for members of the individual and group treatment conditions. Uncontrollable events such as illness and canceled appointments frequently rendered this schedule unattainable. Procedures for clients receiving group and individual treatment did not differ. Group clients were not encouraged to discuss the program, problems, or progress with each other, and opportunities to do so were kept at a minimum. These clients were informed that group treatment permitted us to work with more than one person. Although some of the clients expressed initial concern over the lack of privacy, none refused to participate, and it was not a continuing problem. Each session began with clients completing the Sexual Anxiety Card Sort and providing the previous week's sexual behavior indices. Subjects were then taken into a soundproof room for a videotape session. They were made comfortable in a reclining chair and were instructed to relax while listening to a relaxation tape for approximately 15 minutes. While the sexual scenes

Table 1

Treatment Effects for Initial and 3-Week Follow-up Anxiety, Attitudinal, and Behavioral Measures

Measure	Groups (A)		Time (B)		A × B	
	F	df	F	df	F	df
Anxiety						
Sexual Anxiety Card Sort	6.61**	2, 17	13.12**	6, 108	5.77**	12, 108
Heterosexual Behavior Index	4.30*	2, 17	15.97**	1, 18	6.75**	2, 18
Semantic Differential Anxiety	ns		7.03**	6, 18	2.45**	12, 108
Sexual behavior index						
Female-initiated behavior	ns		2.38*	5, 90	1.93*	10, 90
Visual sexual behavior (male seeing female nude or female seeing male nude)	ns		ns		2.68**	10, 90
Nonsexual sensate focus (male or female giving body massage)	ns		ns		2.37**	10, 90
Sexual sensate focus (the foregoing including breast and genitalia manipulation)	6.80**	2, 18	ns		2.24**	10, 90
Foreplay (cumulative duration 10 minutes)	6.10**	2, 18	ns		ns	
Foreplay (cumulative duration 20 minutes)	6.07**	2, 18	ns		ns	
Intercourse	4.46**	2, 18	ns		5.15**	6, 54
Orgasm	ns		ns		ns	
Sexual concept evaluation						
Sex	11.66**	2, 18	10.02**	6, 108	3.25**	12, 108
Nudity	ns		ns		2.03**	12, 108
Male bodies	ns		3.52**	6, 108	ns	
Self-exploration	ns		ns		2.46**	12, 108
Verbalizing sexual desires	3.75**	2, 18	4.28**	6, 108	ns	
My role in sex	6.41**	2, 18	8.65**	6, 108	3.22**	12, 108
Orgasm	4.05**	2, 18	ns		ns	
Rotter Sex Attitudes	ns		ns		ns	

* $p < .05$.

** $p < .01$.

were on, clients were instructed to visualize both themselves and their partner engaging in the depicted behaviors. During intervals between scenes, they were to utilize whatever relaxation technique they found beneficial. Clients were provided with the option of terminating the scenes by signaling anxiety to the therapist if necessary.

Instruction was provided for the Wolpe (1958) subjective units of distress scale, and distress was to be signaled if the client rated her feelings at greater than 15 units. No client exercised this option. Each of 45 vignettes was presented in brief, with progressively longer portions and relaxation episodes interspersed between them. The 4-min vignettes were presented progressively for 15, 30, 30, 45, 60, and 60 sec with 30-sec pauses between episodes for relaxation. Eight vignettes were shown per session. Viewing sessions lasted approximately 1 hour. After completion, clients filled out the Sexual Semantic Differential. At the end of the session, they were instructed to practice the viewed activities at home with their partners.

At the 3-week follow-up, both partners were interviewed, behavioral indices were collected, questionnaires were readministered, and a standard inquiry into the treatment and personal progress was completed. At the 1-year follow-up, clients were contacted by telephone. All questionnaires, excluding the Rotter Sex Attitude Scale, were readministered.

Subjects assigned to the control condition were told that treatment would not be available for approximately 6 weeks, but they were asked to participate in pretreatment investigations. None of the six women refused, hence all came to the clinic on the same schedule as treatment clients and provided complete accounts of their sexual behavior on the various measures. Male partners in this condition also completed all the sexual behavior indices. Because control group clients were provided with treatment in a different program after time comparable to that required by the treatment groups at the 3-week follow-up interval had elapsed, they were not included in the 1-year follow-up assessment.

Results

The following provides the outcome of analyses on anxiety, behavioral, and attitudinal measures for the duration of treatment and the 3-week and 1-year assessments. Basic analyses of initial treatment effects used a two-way analysis of variance with the two treatment groups and the control condition comprising the three levels of the between-groups factor and the baseline, five process, and 3-week follow-up administrations providing the within-groups factor. Analyses of the 1-year follow-up data used a two-way analysis of variance contrasting the two treatment groups as a between-subjects factor and the four occasions during which data were collected as a repeated measures factor. These were baseline, the fifth treatment session, the 3-week follow-up, and the 1-year follow-up. Significant main effects and interactions were analyzed further using Cicchetti's (1972) version of the Tukey multiple range comparison technique ($\alpha = .05$). Table 1 summarizes the findings for the anxiety, attitudinal, and behavioral measures.

Anxiety Measures

As noted in Table 1, significant main effects and interaction effects for groups over

time were observed for the three anxiety measures. (See Table 2 for mean group frequencies at selected sessions.) Post hoc comparisons on the card sort data indicated that for both treatment conditions, the 3-week follow-up measure differed significantly from the pretest measure as well as from treatment measures at Sessions 1, 2, and 3. There were no differences between measures for the control condition. Post hoc comparisons for the individual assessment sessions indicated that all three groups differed on the pretreatment measure, but the difference disappeared with respect to the individual and group treatment comparison while doubling in magnitude with respect to the differences between both treatment conditions and the control condition. As the premeasures were significantly different, the data were subjected to an analysis of covariance with the premeasure as the covariate. Results indicated significant treatment effects, $F(2, 17) = 4.40, p < .05$, a time effect, $F(5, 85) = 12.81, p < .01$, and a Treatment \times Time interaction, $F(10, 85) = 6.33, p < .01$, indicating that the results were due to treatment over time and not to the initial premeasure difference.

Data for the semantic differential anxiety dimension, obtained by collapsing over the sexual concepts, revealed unfavorable anxiety during the baseline in all groups, with

Table 2
Mean Frequencies for Anxiety Measures

Measure	Baseline	Session 5	3-week follow-up	1-year follow-up
Anxiety Card Sort				
Individual	32.25	12.87	8.00	15.19
Group	22.14	12.42	6.42	9.29
Control	41.66	40.00	45.66	—
Bentler Heterosexual Behavior Hierarchy				
Individual	37.12	—	16.62	17.14
Group	29.28	—	12.00	5.57
Control	40.83	—	46.16	—
Anxiety dimension (Semantic Differential)				
Individual	-13.37	7.50	13.75	14.50
Group	-8.85	7.85	20.14	22.61
Control	-21.00	-18.66	-18.00	—

Note. A dash indicates that the measure was not given at this time.

the treatment groups improving over time until their ratings of the concepts became favorable. The control group initially improved but subsequently declined back to baseline levels. Post hoc analyses indicated that within the two treatment conditions, the follow-up measure differed significantly from the pretreatment and first and second treatment measures. Within each administration period there were no differences between conditions until Sessions 4 through follow-up, at which time the treatment conditions differed from the control condition but not from each other. As with the Sexual Anxiety Card Sort, the decline in anxiety in the treatment conditions manifested itself at the time of the third treatment session.

Findings using the Bentler Heterosexual Behavior Hierarchy were similar, even though this measure was not administered during treatment sessions. For both treatment conditions the follow-up measure was significantly below the baseline measure. The control condition showed no change.

Analyses of the 1-year follow-up data for all three anxiety measures indicated that the two treatment conditions did not differ from each other ($p > .05$), but the effect for the time of administration of the measure was significant. This was the case for the card sort, $F(3, 12) = 25.20$, $p < .01$, the semantic differential, $F(3, 12) = 9.38$, $p < .01$, and the Bentler hierarchy, $F(3, 12) = 9.39$, $p < .01$. Post hoc comparisons indicated that the 1-year follow-up data differed significantly from the baseline measure but not from treatment termination or the 3-week follow-up. Thus, the reduction in anxiety displayed at treatment termination and the 3-week follow-up was maintained.

Behavioral Measures

Clients' and partners' reports for the different categories of the sexual behavior index were intercorrelated to evaluate reliability. Following z-score transformations, an overall mean intercorrelation of .912 was obtained, indicating good reliability. The following provides details on the clients' data only.

Treatment influenced the extent to which

the women initiated sexual behavior with their partners, as indicated by the significant Group \times Time interaction. Consistent with the means reported in Table 3, post hoc pairwise comparisons indicated that clients in the two treatment conditions reported increases in the number of female initiatives, whereas the no-treatment control condition actually declined over the 6 weeks of monitoring. These effects were maintained throughout the year preceding the 1-year follow-up, as indicated by a main effect for the four measurement periods, $F(3, 12) = 6.55$, $p < .01$, without differences between the 3-week and 1-year follow-up data. The two treatment groups did not differ at the 1-year follow-up.

Treatment effects over time were also observed for three of the five measures of pre-coital sexual interaction. With respect to visual sexual behavior, post hoc comparisons of the significant Treatment \times Time interaction indicated that clients in the two treatment conditions maintained stable rates of visual exposure to their partners, and the no-treatment control groups actually declined in the frequency of visual exposure over the 6 weeks of behavioral monitoring.

Analyses of long-term effects for visual sexual exposure revealed that neither of the two treatment groups differed from the rate of visual exposure observed at the 3-week follow-up. However, the two groups did differ significantly from each other at that time, with the group treatment exceeding the individual treatment, $F(1, 12) = 8.25$, $p < .01$.

Nonsexual sensate focus effects were apparently the result of those clients receiving group treatment engaging in significantly more nonsexual sensate focus, relative to the baseline measure, whereas clients in the no-treatment control actually experienced a significant decrease in the monitored behavior. Clients in the individual treatment condition maintained a fairly steady rate of nonsexual sensate focus.

The results of the 1-year follow-up for nonsexual sensate focus indicated that the two treatment groups did not display any significant change over the follow-up period ($p > .10$).

Post hoc comparisons indicated that both

Table 3
Mean Frequencies for Behavioral Index

Measure	Baseline	Week 2	3-week follow-up	1-year follow-up
No. female initiatives				
Individual	.87	2.62	2.39	1.85
Group	.57	1.42	2.30	2.80
Control	1.33	.66	.50	—
Visual sexual exposure				
Individual	10.12	8.25	7.62	10.42
Group	9.14	13.14	12.42	16.00
Control	12.83	6.83	3.66	—
Nonsexual sensate focus				
Individual	1.62	1.37	1.50	3.71
Group	.42	3.85	3.14	3.57
Control	3.50	1.83	.53	—
Sexual sensate focus				
Individual	6.00	11.62	8.45	9.71
Group	4.71	9.71	10.57	12.71
Control	4.50	2.16	.75	—
Foreplay duration 10 min.				
Individual	.62	1.75	1.13	1.57
Group	2.00	1.42	1.34	1.57
Control	.16	.00	.16	—
Foreplay duration 20 min.				
Individual	.37	1.25	.87	1.28
Group	.28	.42	.76	1.50
Control	.33	.00	.10	—
Intercourse				
Individual	1.62	2.37	2.31	2.14
Group	1.00	1.57	2.14	3.85
Control	1.16	.66	.50	—
Orgasm				
Individual	.50	.62	.37	1.42
Group	.00	.28	.42	1.00
Control	.00	.00	.00	—

Note. A dash indicates that the measure was not given at this time.

treatment groups displayed initial increases in the frequency of sexual sensate focus that were sustained throughout treatment and during the first week of follow-up. However, the individual treatment clients declined thereafter relative to the group treatment clients. The control group maintained a relatively stable incidence of sexual sensate focus throughout the 6 weeks of self-monitoring, and, at the time of the 3-week follow-up, they reported significantly lower levels than either of the two treatment groups. The in-

creased incidence of sexual sensate focus was maintained from the 3-week to the 1-year follow-up, with the two treatment groups not changing over this interval ($p > .10$) and not differing from one another at that time.

Analyses of clients' reports of the incidence of foreplay lasting longer than 1 minute but less than 10 minutes indicated a significant main effect for treatment, with subsequent post hoc comparisons indicating that the two treatment groups differed significantly from the control group but not from each other.

Inspection of the data indicates a very high incidence of 10 minutes of foreplay in the group treatment clients' reports during the baseline period. At the time of the 1-year follow-up, the two treatment groups maintained the incidence reported at the 3-week follow-up.

Analyses of the incidence of foreplay lasting between 10 and 20 minutes also indicated significant treatment effects but no interaction with time. Paired comparisons of the main effect indicated that the two treatment groups reported a greater incidence than the control group, but they did not differ from one another. No initial differences were observable in the baseline measure. There were no differences between the 3-week follow-up and the 1-year follow-up, indicating that the beneficial effects of treatment had been maintained.

Analyses of the incidence of intercourse were restricted to differences between the baseline and 3-week follow-up measures because clients and their partners had been enjoined to avoid intercourse throughout treatment. Results indicated significant treatment and Treatment \times Time interaction effects. Post hoc comparisons indicated that the significant Time \times Treatment interaction apparently was the result of an increased frequency of intercourse relative to baseline for those clients receiving group and individual treatment as compared to a decline in incidence as reported by the no-treatment controls. At the time of the 1-year follow-up, post hoc analyses of a significant interaction between groups and time, $F(3, 12) = 10.77$, $p < .01$, indicated that the group receiving individual treatment maintained the level of intercourse observed at the time of the 3-week follow-up. However, group treatment led to significant increases relative to both the individual treatment group and the incidence of intercourse reported at the time of the 3-week follow-up.

Analyses of the incidence of orgasm reported by the women yielded no significant effects for treatment or the interaction of time with the treatment condition. Reports of the incidence of orgasm were at the zero level for the control group and were only marginally greater for either of the two treat-

ment groups. At the end of the 1-year follow-up, there was a nonsignificant tendency for both treatment groups to increase in the incidence of orgasm relative to the baseline, treatment termination, and 3-week measures, $F(3, 12) = 1.23$, $p > .10$.

Attitudinal Measures

The Rotter Sex Attitude Scale was administered to evaluate global sexual attitudes, and a semantic differential had been constructed to examine the evaluative dimension with respect to specific sexual concepts. Data from the Adjusted-Maladjusted subscale of the Rotter measure were subjected to a two-way analysis of variance, with the treatment groups and control condition comprising the three levels of a between-groups factor and the baseline, final session, and follow-up administrations providing the within-groups factor. The analysis revealed no significant main effects or interactions. Mean values for all groups indicated adequate sexual adjustment.

The Sexual Semantic Differential scores indicated the degree of positive or negative affect induced in the client by each of the seven concepts relating to either specific sexual activities or concepts with sexual arousal potential. In contrast to the Rotter scale, attitudinal changes and treatment effects were observed, as summarized in Table 1. Interaction effects were observed between treatment and time for the following sexual concepts: "sex," "nudity," "self-exploration," and "my role in sex." Means for groups at the different times at which the measures were administered appear in Table 4. With respect to the concept *sex*, post hoc comparisons indicated that whereas all groups initially had unfavorable attitudes, the two treatment groups progressively developed favorable attitudes through the 3-week follow-up, and the control condition remained at the baseline level throughout. The concept *nudity* was met with an essentially neutral response in all groups at the time of the baseline measure. The control group did not change over time, but both treatment groups exhibited favorable attitudes at the time of the 3-week fol-

Table 4
Mean Frequencies for Semantic Differential

Sexual concept	Baseline	Session 5	3-week follow-up	1-year follow-up
Sex				
Individual	-6.28	7.14	10.71	13.00
Group	-5.28	10.14	10.85	13.85
Control	-13.16	-12.00	-12.50	
Nudity				
Individual	5.14	8.10	11.57	11.14
Group	8.42	8.10	13.57	13.71
Control	3.50	.00	-.50	—
Male bodies				
Individual	4.85	8.42	11.14	9.85
Group	6.14	10.71	14.85	14.14
Control	1.16	1.00	.50	—
Self-exploration				
Individual	2.42	5.14	8.75	6.42
Group	4.42	6.28	9.71	6.71
Control	2.83	-3.50	-4.00	—
Verbalizing sexual desires				
Individual	-.28	2.28	5.14	3.85
Group	-4.10	6.00	8.14	7.42
Control	-13.50	-8.50	-10.00	—
My role in sex				
Individual	-11.57	5.80	9.10	10.85
Group	-3.28	6.00	9.80	12.00
Control	-15.83	-11.50	-13.66	—
Orgasm				
Individual	.12	9.25	11.25	18.57
Group	10.71	11.57	14.00	13.00
Control	.50	-4.00	-1.50	—

Note. A dash indicates that the measure was not given at this time.

low-up. The concept *self-exploration* initially yielded mildly favorable responses in all groups, with the two treatment groups subsequently improving and the control group deteriorating marginally through the 3-week follow-up. The concept *my role in sexual activity* was initially met with negative attitudes in all groups, with the two treatment groups ultimately providing evidence of favorable attitudes and the control group manifesting no change.

The concept *verbalizing sexual desires* yielded significant treatment effects and changes over time. As indicated in Table 4, the group differences can be attributed to much more favorable attitudes in the two treatment

groups than in the control group. However, the control group differed substantially from the two treatment groups at the time of the baseline measure, and it would be inappropriate to attribute differences to treatment effects. Similarly, all groups apparently improved in their evaluation of the concept over time, but this again could not be attributed to treatment because the control group improved as well. It is noteworthy that the treatment groups ultimately ended up with favorable attitudes while the control group maintained an unfavorable evaluation of the concept.

Analyses of evaluative attitudes toward the concept *orgasm* indicated that the groups

differed significantly, but there were no changes over time nor did group effects interact with time. At the time of the baseline measure, those individuals about to receive group treatment appeared to have substantially more favorable attitudes than those about to receive individual treatment or those in the control group. The two latter groups were essentially neutral with respect to their attitude toward the concept. At the time of treatment termination and the 3-week follow-up, the control group maintained the essentially neutral attitude, with the two treatment groups quite positive and not different from one another. The three groups could not be differentiated in terms of their attitudes toward the concept *male bodies* at any time during baseline, treatment, or follow-up. However, attitudes tended to become more favorable over time in all three groups.

Analyses of the 1-year follow-up data indicated that for the majority of the sexual concepts (sex, nudity, my role in sexual activity, verbalizing sexual desires, and orgasm), treatment gains observable at the time of the 3-week follow-up were maintained. No significant differences were observable between measures taken on the two occasions ($p > .05$). In addition, there were no significant differences between groups receiving individual or group treatment after 1 year. With respect to the concept *self-exploration*, there was a nonsignificant decline in attitude through the 1-year follow-up, suggesting that after this time early benefits from self-exploration had led to other more satisfying forms of experience. No change appeared in characterizations of the concept *male bodies*.

Discussion

Clients receiving group and individual graduated symbolic modeling with concurrent behavioral tasks improved significantly with respect to sexual anxiety, attitudes, and behavioral enactment, in contrast to the no-treatment control condition. Although no statistically significant differences were observed between group and individual treatment programs, clients receiving group treatment recorded less anxiety on the card sort and the anxiety dimension of the Sexual Semantic Dif-

ferential. There was also a tendency for clients receiving the group treatment to report favorably higher frequencies on the sexual behavior index. As clients were randomly assigned to the two treatment conditions, there was no reason to believe that those receiving group treatment were different from those receiving individual treatment. It is possible that components of the group treatment, such as exposure to people with similar problems, induced lower levels of anxiety and facilitated the incidence of sexual activity.

There was also a tendency toward deterioration in the control group over all indices. Behavioral frequencies generally tended to decline, and in two cases, visual sexual exposure and nonsexual sensate focus, they decreased significantly. On several other measures, there was a pattern of initial improvement with subsequent return-to-baseline incidence. There appeared to be an initial enthusiasm for involvement in what was described at the time as a "pretreatment investigation," which essentially comprised self-monitoring over the span of about 6 weeks; however, the improvement waned with time. It seems that the demand-induced increased awareness of personal sexual habits provoked anxiety and dissatisfaction to the extent that the initial favorable effects of self-monitoring (Kanfer, 1970; Kazdin, 1974) declined rapidly after the third session. It would be useful to know if clients on waiting lists for sexual dysfunction treatment all manifest this trend toward deterioration.

Because improvement was observed around the third treatment session on all measures of behavior, attitude, and anxiety, it appears that the treatment effects were consolidated simultaneously in both covert and overt behavioral systems. Including behavioral measures throughout the course of treatment not only permitted assessment of when therapeutic effects manifested themselves but indicated whether clients were complying with treatment instructions.

Results from the several measures of anxiety, the Sexual Anxiety Card Sort, the Bentler Heterosexual Behavior Hierarchy, and the Sexual Semantic Differential all indicated that both treatment programs reduced anx-

iety throughout treatment, with reduced anxiety persisting through both the 3-week and 1-year follow-up. The control subjects displayed a nonsignificant trend toward increased anxiety through the 3-week follow-up. Reductions in reported anxiety as measured by the Bentler were slightly less than those observed on the Sexual Anxiety Card Sort. As the Bentler describes more novel sexual activities than the Sexual Anxiety Card Sort, there was evidence of generalized treatment effects.

Because the Marks and Sartorius (1968) Sexual Semantic Differential was originally devised to assess sexual attitudes toward unconventional sexual activities, it is noteworthy that it was of value in the present context, thereby attesting to its flexibility and usefulness. In the initial theoretical formulation, it was expected that positive increases on the Sexual Semantic Differential would be more marked for those concepts conforming to the behavioral treatment requirements, namely, sex, self-exploration, verbalizing sexual desires, and my role in sex. The findings were partially consistent with this expectation, since treatment effects were particularly potent for the concepts *sex* and *my role in sexual activity*. Significant differences between the two treatment groups and the control group generally manifested themselves after the third treatment session. These results have implications for time-limited therapy in that treatment effects appear to consolidate themselves within the attitudinal realm fairly early during treatment programs.

In accordance with expectations, no significant changes were evidenced over the three administrations of the Rotter Sex Attitude Scale. It did not seem probable that the treatment would influence the more global attitudes measured by this scale. It was interesting to note that none of the clients scored within the maladjusted range on this scale, suggesting that personal sexual dysfunction may be entirely separate from global sexual attitudes.

The sexual behavior index disclosed treatment-induced improvement in the categories measuring the incidence of visual sexual exposure, nonsexual sensate focus, sexual sensate

focus, the number of female-initiated sexual behaviors, and the frequency of intercourse. These effects were maintained without exception through the 1-year follow-up. Although an increased incidence of orgasm tended to be a personal goal for most of the clients, it was not affected by treatment, a finding consistent with Wincze and Caird (1976). They agree with Kaplan (1974), who observed that orgasms can be conceptualized as partially independent of sexual pleasure. As such, achieving orgasm is the appropriate treatment objective of complementary, but different, therapeutic programs.

Although no statistically significant differences were observed between group and individual treatment conditions, there were tendencies for those clients receiving group treatment to engage in more visual sexual exposure, nonsexual sensate focus, and female-initiated sexual behaviors. When debriefed, the women receiving group treatment tended to express positive feelings concerning it. The major issues included a feeling of group cohesion, an increased ability to describe personal sexual experiences in public, and a sense that they wished to do well for other members of the group. Given that group treatment is far more economical, this type of program appears to have some decided advantages. The above findings indicate the viability of the group treatment as a highly desirable alternative to conventional individual treatment.

Reference Notes

1. Barlow, D., Leitenberg, H., & Agras, S. *The experimental control of sexual deviation through manipulation of the noxious scenes in covert sensitization*. Paper presented at the meeting of the Eastern Psychological Association, Washington, D.C., April 1968.
2. Rotter, G. S. *Sex Attitude Scale*. Unpublished, 1973.

References

- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Bandura, A., Blanchard, E., & Ritter, B. Relative efficacy of desensitization and modeling approaches for inducing behavioral, affective and attitudinal changes. *Journal of Personality and Social Psychology*, 1969, 13, 173-199.

- Bandura, A., Grusec, J. E., & Menlove, F. L. Vicarious extinction of avoidance behavior. *Journal of Personality and Social Psychology*, 1967, 5, 16-23.
- Bentler, P. Heterosexual behavior assessment: II. Females. *Behaviour Research and Therapy*, 1968, 6, 27-30.
- Brady, J. Brevital relaxation treatment of frigidity. *Behaviour Research and Therapy*, 1966, 4, 71-77.
- Cicchetti, D. V. Extension of multiple-range tests to interaction tables in the analysis of variance: A rapid approximate solution. *Psychological Bulletin*, 1972, 77, 405-408.
- Clopton, J. R., & Risbrough, R. F. Sexual arousal for desensitization. *Behavior Therapy*, 1973, 4, 741-743.
- Craig, K. D., Best, H., & Ward, L. M. Social modeling influences on psychophysical judgments of electrical stimulation. *Journal of Abnormal Psychology*, 1975, 84, 366-373.
- Kanfer, F. H. Self-monitoring. *Journal of Consulting and Clinical Psychology*, 1970, 35, 148-152.
- Kaplan, H. *The new sex therapy: Active treatment of sexual dysfunction*. New York: Brunner/Mazel, 1974.
- Kazdin, A. E. Self-monitoring and behavior change. In M. J. Mahoney & C. E. Thoresen (Eds.), *Self-control: Power to the person*. New York: Holt, Rinehart & Winston, 1974.
- Kraft, T., & Al-Issa, I. A. Behavior therapy and the treatment of frigidity. *American Journal of Psychological Therapy*, 1967, 21, 116-120.
- Lazarus, A. The treatment of chronic frigidity by systematic desensitization. *Journal of Nervous and Mental Disease*, 1963, 136, 272-278.
- Marks, I. M., & Sartorius, N. H. A contribution to the measurement of sexual attitude. *Journal of Nervous and Mental Disease*, 1968, 115, 441-451.
- Masters, W. H., & Johnson, V. E. *Human sexual inadequacy*. Boston: Little, Brown, 1970.
- Wincze, J. P., & Caird, W. K. The effects of systematic desensitization and video desensitization in the treatment of essential sexual dysfunction in women. *Behavior Therapy*, 1976, 7, 335-342.
- Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford, Calif.: Stanford University Press, 1958.
- Woody, R., & Schauble, P. Videotaped vicarious desensitization. *Journal of Nervous and Mental Disease*, 1969, 148, 281-286.

Received October 4, 1976 ■

Behavioral Treatment of Problem Drinkers: A Comparative Outcome Study of Three Controlled Drinking Therapies

William R. Miller
University of New Mexico

Three therapies designed to reduce the alcohol consumption of problem drinkers were evaluated and compared. Self-referred and court-referred clients were randomly assigned to one of three treatment groups: aversive counterconditioning (AC) using self-administered electrical stimulation; behavioral self-control training (BT) including self-monitoring and instruction in functional analysis; or a controlled drinking composite (CD) including blood alcohol awareness training, discriminated aversive counterconditioning, self-monitoring, and rate-control training. All therapies were conducted by trained paraprofessionals and consisted of 10 weekly sessions. All three behavior therapies produced significant reduction in clients' weekly alcohol consumption and peak blood alcohol concentration. These gains were largely maintained over 1 year of follow-up. No significant differences among treatments were found, although AC was initially least effective in reducing alcohol consumption and producing improvement on measures of general adjustment. CD proved least cost effective, requiring four times more therapist contact than either AC or BT. A self-control manual was found to be helpful in maintaining controlled drinking.

The importance of alcohol abuse as a national health problem is apparent. Less obvious is its appropriate remedy. Adherents of the traditional disease model of alcoholism have maintained that it involves an irreversible loss of control over drinking and that the resumption of normal drinking is impossible for the alcoholic (Alcoholics Anonymous, 1955; Jellinek, 1960). The treatment of problem drinkers, dominated for the past 30 years by this viewpoint, has focused primarily on the goal of total and lifelong ab-

stinence. The general applicability of this model has been increasingly questioned, however (Armor, Polich, & Stambul, 1976; Miller & Caddy, 1977; Sobell & Sobell, 1974), and treatment methods designed to produce controlled drinking have begun to emerge (e.g., Lovibond & Caddy, 1970; Miller & Muñoz, 1976; Sobell & Sobell, 1973b).

The status of these controlled drinking therapies is uncertain at present (Nathan, 1977). A wide variety of treatment methods have been proposed for the reduction and control of drinking (Briddell & Nathan, 1975; Hamburg, 1975; Lloyd & Salzberg, 1975; P. M. Miller, 1976; Nathan, 1976). W. R. Miller (1976) has reviewed outcome research regarding seven major strategies: (a) aversive counterconditioning, (b) blood alcohol discrimination training, (c) rate-control training, (d) operant methods, (e) stimulus control procedures, (f) self-monitoring, and (g) broad-spectrum approaches. Although these techniques have been within the armamentarium of behavior therapists for some time, their *relative* efficacy in the control of drinking has not been established. To further

The present article is based on dissertation research conducted at the University of Oregon. Special appreciation is expressed to the members of my committee: Edward Lichtenstein, Robyn M. Dawes, Susan K. Gilmore, Lewis R. Goldberg, Neil McNaughton, and Lonnie Snowden.

This research was supported in part by a grant from the Distilled Spirits Council of the United States, Inc. Supplemental support was provided by the Christian Brothers and Hillcrest Vineyards and by the Department of Psychology, University of Oregon.

Requests for reprints should be sent to William R. Miller, Department of Psychology, University of New Mexico, Albuquerque, New Mexico 87131.

complicate matters, outcome studies have used varying client populations (albeit primarily inpatient alcoholics) and have, with few exceptions (e.g., Sobell & Sobell, 1973a, 1976), failed to provide long-range follow-up data.

The present study was designed to assess the relative effectiveness within an outpatient setting of three controlled drinking therapies: (a) aversive counterconditioning, (b) behavioral self-control training, and (c) direct training in rate control and blood alcohol discrimination. On the basis of the existing literature (W. R. Miller, 1976), it was predicted that all three therapies would produce significant reduction in drinking behavior over the course of treatment. Of interest was the relative effectiveness of the three methods and specifically whether the treatments would differ with regard to (a) their impact on drinking behavior, (b) their impact on measures of psychological functioning, and (c) the maintenance of whatever behavior changes they produced. Previous research has provided no basis on which to predict the relative effectiveness of the three methods.

Method

Clients

A treatment program for problem drinkers who desired "to reduce and control their drinking without stopping altogether" was announced through the local news media and mental health service agencies of Eugene, Oregon. Clients came to the program through either of two routes: self-referral or court referral.

Self-referrals. Of 45 persons voluntarily seeking treatment, 1 was excluded due to advanced liver disease. Nine others dropped out of the program prior to or during the initial assessment phase and before beginning treatment. An additional 5 clients dropped out of treatment prior to the third session, and 1 client terminated following the eighth session. The remaining 29 clients completed the 10-session treatment program.

Court referrals. Oregon law provides for the alternative sentencing to a treatment program of persons found guilty of driving while intoxicated. The sons found guilty of driving while intoxicated. The Alcohol Traffic Safety Clinic (ATSC), responsible for treating such offenders, referred selected clients to this program. During the period of intake for the present study, a total of 89 clients were processed by the ATSC.

To enter the program a court-referred client had to pass through several screening stages. Clients unwilling to sign a release permitting evaluation of

their progress ($n=10$) were excluded altogether. An additional 8 clients were incarcerated or left the community before they could be evaluated at intake. These 18 clients were thus not studied.

The ATSC staff further excluded 35 clients as inappropriate for controlled drinking. The remaining 36 clients were considered eligible for behavioral treatment. Of these, 6 refused or dropped out prior to treatment. Four began treatment but dropped out before the third session, and 1 terminated after the fourth session. All court-referred clients who dropped out of the program were referred back to ATSC for alternative therapies (primarily group and/or Antabuse). Eight clients who were otherwise eligible for behavioral treatment were randomly selected to be assigned back to ATSC as a "random" control group. A final total of 17 clients signed the release, were cleared by ATSC staff, were randomly assigned to and accepted treatment, and completed the 10 sessions.

Client Assignment Procedures

Clients were randomly assigned to one of the three treatment modes and to 1 of 10 paraprofessional therapists. All clients were informed prior to assessment and treatment of the modality to which they had been assigned. The nature of the treatment, including its potential risks and discomforts, was fully explained, and all clients read and signed a statement of informed consent for their particular treatment.

Project Staff

All treatment sessions were conducted by paraprofessional therapists trained in the behavioral procedures described below. All therapists treated clients within each of the three treatment modalities.

Therapists were recruited from among graduate and undergraduate students in psychology and allied fields and received academic credit for their participation. Eight therapists were women and two were men. By group means, the "average" therapist was 26.5 years of age, had completed 4.5 years of postsecondary education, and had had 3 years of prior experience in some form of human services. Therapists followed a detailed procedures manual for treatment sessions (Miller, Note 1) and received ongoing supervision.

All assessment procedures were administered by trained interviewers who had no contact with or responsibility for therapy. Three different teams of interviewers conducted assessment at intake and termination, at 3-month and 1-year follow-ups.

Facilities and Apparatus

A conventional interview room served as the setting for assessment and for the treatment of clients in the behavioral self-control training group. A simulated bar setting provided the environment for

the other two treatment modalities, with the therapist in the usual position of a bartender. An Omicron Intoxilyzer, Model 4011, was installed in the bar to provide feedback regarding blood alcohol concentration. Equipment for the administration of electric shock included a Lafayette Master Shocker (Model 82400) with an intensity range of 0 to 5 ma, two Lafayette Interval Timers (Model 58010), and a Lafayette Probability Generator (Model 58019). A triple-function data recorder (Lafayette Model 58004) was installed on the bar top as an in-session counter for shocks, sips, and other relevant events. A quad indicator panel (Lafayette Model 58001) was connected with the interval timers to provide signals for the proper sequencing and spacing of aversive conditioning trials.

Initial Assessment

All clients completed 4 hours of assessment prior to behavioral treatment. The first 2-hour session included a demographic questionnaire, the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943), the Michigan Alcoholism Screening Test (MAST; Selzer, 1971), Rotter's Internal-External Locus of Control (I-E) Scale (Rotter, 1966), and a Profile of Mood States (McNair, Lorr, & Droppelman, 1971). The second 2-hour session consisted of the Drinking Profile (Marlatt, 1976), a structured interview about drinking patterns and habits. Written consent was obtained from each client to contact at least two significant others for corroborative information. Whenever possible, parallel intake assessment (with the exception of the Drinking Profile) was conducted with clients who dropped out of or were excluded from the program.

Treatments

Aversive counterconditioning. The aversive counterconditioning (AC) program took place in the simulated bar setting and required 10 weekly 30-minute sessions. The client was seated at the bar and attached two plate electrodes to the ring and index fingers of his or her nondominant hand. Shock intensity, which the client could change at any time, was set at a level that he or she had rated to be painful but not intolerable. The client's one or two most frequently consumed beverages were prepared and placed on the bar.

A single trial consisted of several steps. The client was instructed to pick up the beverage, sniff it, and imagine tasting it without actually sipping. The client then pressed a foot pedal installed near the bar stool, administering a 1-sec shock. Shocks occurred only 70% of the time, being randomly scheduled by the probability generator. Upon pressing the pedal, the client put the drink back on the bar and removed his or her hand from it. The importance of timing in these steps was emphasized, and the therapist closely observed the trials. A timing system sustained a "hold" light and deactivated the

shock button for 15 sec, after which a "go" light signaled the beginning of the next trial. Between trials the client was allowed to adjust shock intensity if desired, being encouraged to increase intensity over trials. An AC session was terminated after a minimum of 40 or a maximum of 50 trials.

Clients in all treatment modalities were trained in self-monitoring and kept daily record cards of all drinks during each week of treatment. Clients in AC submitted these records to the therapist during weekly sessions, but no specific use was made of them in treatment.

Behavioral self-control training. All 10 behavioral self-control training (BT) sessions were conducted in the interview room and lasted approximately 30 minutes each. As in the other two treatment modalities, clients began self-monitoring after the first session. The record cards kept by clients included columns for recording the type and amount of each drink consumed, time of the first sip, and situational information. These cards served as the focus of BT sessions.

During each session the client reviewed the records of her or his previous week's drinking. The therapist completed basic calculations regarding alcohol consumption and blood alcohol concentration and helped the client to examine the records for stimulus antecedents of heavy drinking. Strategies for increasing future control were discussed. Suggested self-control methods encompassed three major strategies: (a) awareness and alteration of antecedents of drinking, (b) reduction of drinking rate, and (c) identification and practice of alternatives to the use of alcohol. These strategies have been outlined by Miller and Muñoz (1976).

Controlled drinking composite. The simulated bar setting served as the setting for controlled drinking (CD) sessions, which lasted between 120 and 150 minutes each. Because clients consumed alcohol during these sessions, arrangements were made for transportation to and from the clinic.

During Session 1 the client was provided with her or his most frequently consumed beverages and was instructed to drink as usual. No client was permitted during any session to consume drinks that would elevate the blood alcohol concentration above 80 mg%. Training in self-monitoring was included in this first session.

During Sessions 2-4 the client was instructed to drink in his or her normal manner and was taught to estimate the blood alcohol concentration reached after each drink by attending to both internal and external cues.¹ When the client's blood alcohol con-

¹ Sessions 1-5 had originally been intended to include blood alcohol concentration discrimination training after the manner of Lovibond and Caddy (1970). Malfunctioning of the Intoxilyzer, however, precluded its use for this purpose. When it became apparent that the malfunction could not be corrected, a procedure was substituted whereby clients were taught to estimate their own blood alcohol

centration reached 65 mg% or 30 minutes before the end of a session (whichever came first), electrodes were attached as in AC, and discriminated aversive counterconditioning (Lovibond & Caddy, 1970) was begun. During Session 2, 10 shocks were self-administered and were paired with sips of a beverage. During Sessions 3-4, 40 trials were required, pairing shocks with prepping behaviors as in AC. Conversation between therapist and client during these and subsequent sessions focused on drinking behavior, antecedents and consequences of alcohol use, alcohol education, and self-control strategies (Miller & Muñoz, 1976). Daily record cards were examined during each session to facilitate discussion of the previous week's drinking pattern.

Session 5 served as a probe session. No shocks were administered, but the client continued to record drinks, sips, and blood alcohol concentration estimates.

Sessions 6-8 focused on the reduction of alcohol consumption rate (Sobell & Sobell, 1973b). Electrodes were attached as before, but shocks were controlled by the therapist during these sessions. A list of rules similar to those used by Sobell and Sobell (1973b) was read to the client, indicating the behaviors for which a shock would be delivered. Shock intensity was set at the highest level previously tolerated by the client.

Session 9 served as a final probe, again with no shock contingency in effect. The therapist observed rule infractions but did not punish them.

Session 10 consisted of a review of the treatment process and was conducted in the interview room. A thorough functional analysis was constructed from information gained during treatment, and in vivo applications of self-control procedures were reviewed.

Termination Assessment

A portion of the final session in each treatment mode was devoted to the completion of several assessment forms, including the Profile of Mood States and ratings of treatment and therapist satisfaction. Significant others were contacted to assess current drinking behavior. Clients were asked to keep daily record cards for an additional 12 weeks, mailing them weekly. Whenever possible, parallel information was obtained from clients not receiving behavior therapies.

One half of the clients in each of the three treatment groups were randomly chosen to receive a

copy of a manual prepared as an aid in the maintenance of treatment gains (Miller & Muñoz, Note 2).

Follow-up Assessment

At 3 months following termination, clients were contacted and invited to participate in a follow-up session. A fee of \$5 was paid to each client for participation. Instruments administered at this time included a questionnaire paralleling previous measures of drinking behavior, the MMPI, Rotter's I-E scale, the Profile of Mood States, and (where appropriate) a questionnaire regarding the client's use of the manual. Significant others were again contacted for corroborative information. Driving records were obtained for all court-referred clients. Those who had not previously received a copy of the manual were given one at this time.

Whenever possible, clients not treated in the program were also interviewed and paid for their participation. To determine the appropriate time for these follow-up interviews, all nontreated clients were yoked to treated clients according to the date of their intake with the program or with the ATSC.

At 1 year following treatment termination, clients and significant others were once again interviewed by telephone and mail. Drinking behavior was assessed, and clients again completed the Profile of Mood States.

Results

Pretreatment Measures

Table 1 presents demographic data for each of the six groups studied. The groups differed on several important pretreatment measures. Contrasting self-referred with court-referred clients receiving behavioral treatment, the former reported significantly higher income and education and scored significantly higher on the MAST and on three subscales of the MMPI: Depression, Psychopathic deviate, and Rich and Davis' (1969) Revised Alcoholism Scale. Self-referred clients also reported a significantly higher rate of alcohol consumption at intake than did court-referred clients, $F(1, 40) = 19.67, p < .001$. No significant differences on pretreatment measures were found among the three behavioral treatment groups, consistent with random assignment.

Within the population of traffic violators, clients excluded by the ATSC were found to be significantly older and less educated and to have scored significantly higher on the MAST and on the three MMPI subscales mentioned above.

concentration from number of drinks consumed, body weight, and spacing of drinks. Other research has indicated this to be an equally effective procedure for blood alcohol concentration training (Huber, Karlin, & Nathan, 1976). This change necessitated the repetition of one or two sessions for five CD clients who had been using the Intoxilyzer.

Table 1
Demographic Information Regarding Client Groups

Client group	■		Age	Years of education	Annual income
	Males	Females			
Self referred					
Completed treatment	18	11	37.3	15.6	\$14,703
Refused or dropped out	11	4	32.9	14.1	\$10,034
Court referred					
Completed treatment	14	3	35.8	11.9	\$ 8,227
Excluded from treatment	30	5	41.5	11.3	\$8,384
Refused or dropped out	9	2	27.3	12.4	\$ 8,305
Random controls	6	2	41.3	13.5	\$ 9,955

Effects of Treatment on Drinking Behavior

Self-report. Clients were interviewed at intake, termination, and at 3- and 12-month follow-ups. Reported alcohol consumption at each point was converted into standard ethanol content units, with 1 standard ethanol content unit equal to .5 ounce (15 ml) of pure ethanol. Figure 1 shows the mean self-reported alcohol consumption of each treated group at each of the four assessment periods. A repeated measures analysis of variance revealed a significant decrease in drink-

ing for all groups, $F(3, 81) = 12.53, p < .001$. No significant differences among groups were observed.

Reports of significant others. Mean estimates of significant others regarding clients' drinking are presented in Figure 2. In cases in which multiple estimates were obtained, the highest estimate was used. Product-moment correlations between the reports of clients and of their significant others were calculated and yielded a puzzling pattern. At intake there was no significant relationship

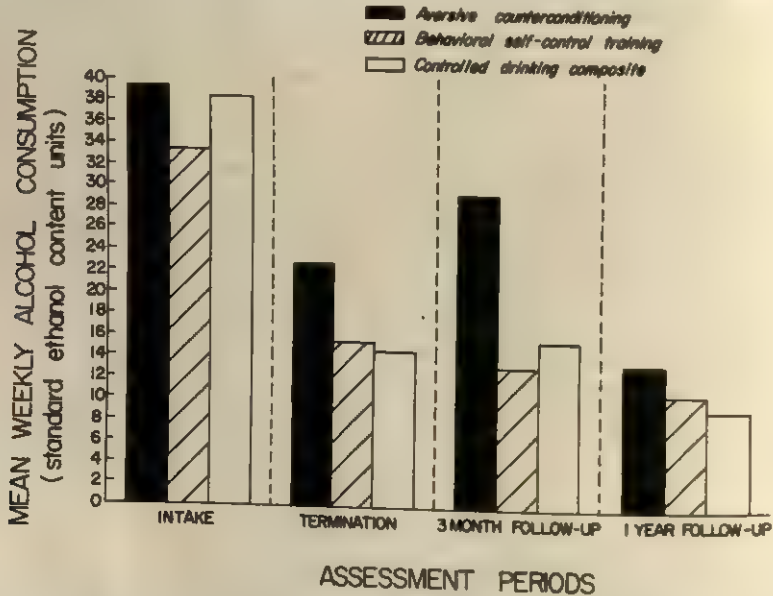


Figure 1. Mean self-report of weekly alcohol consumption. (From Chapter 6 "Behavioral Self-control Training in the Treatment of Problem Drinkers" by William R. Miller. In *Behavioral Self-management*, Richard B. Stuart (Ed.). Copyright by Brunner/Mazel (1977). Reprinted by permission.)

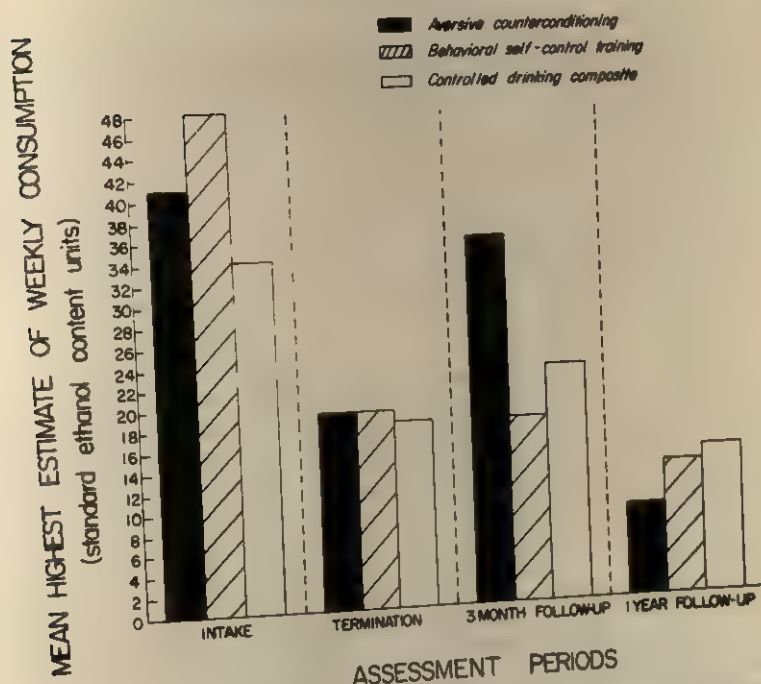


Figure 2. Mean highest estimate from significant others of clients' weekly alcohol consumption. (From Chapter 6 "Behavioral Self-control Training in the Treatment of Problem Drinkers" by William R. Miller. In *Behavioral Self-management*, Richard B. Stuart (Ed.). Copyright by Brunner/Mazel (1977). Reprinted by permission.)

between these two data sources, $r(28) = .225$, $p > .05$. At termination there was a mild positive relationship, $r(28) = .365$, $p < .05$, whereas at the 3-month, $r(29) = .733$, $p < .001$, and 12-month follow-ups, $r(20) = .821$, $p < .001$, the reports were highly correlated. At each assessment period an approximately equal number of significant others overestimated and underestimated clients' own reports of their drinking.

Daily record cards. Alcohol consumption and weekly peak blood alcohol concentration as reported on record cards over the weeks of treatment and follow-up are presented in Figures 3 and 4, respectively. Repeated measures analyses of variance indicated significant decreases over the course of treatment both in weekly consumption, $F(8, 344) = 8.51$, $p < .001$, and in weekly peak blood alcohol concentration, $F(8, 344) = 4.52$, $p < .001$. Blood alcohol concentration peaks were estimated from body weight and consumption data (Miller & Muñoz, 1976). There were again no significant differences among treatment

modalities. In time series analyses of these data (Glass, Willson, & Gottman, 1972), all three groups showed a significant downward drift in consumption during treatment: for aversive counterconditioning, $t(17) = -3.93$, $p < .001$; for behavior training, $t(17) = -8.77$, $p < .001$; and for controlled drinking, $t(17) = -13.30$, $p < .001$, and maintained these gains, with level slopes during follow-up.

Improvement ratings. The use of improvement ratings provides an index of the status of individuals following treatment. Each client was assigned to one of six improvement categories at termination and at 3- and 12-month follow-ups.

Because a number of clients (primarily court referred) were found to be drinking very little at intake, an initial distinction was drawn between those who had been "problem drinkers" and those who had been "controlled drinkers" immediately prior to treatment. A problem drinker was defined as anyone who exceeded 20 standard ethanol content units per week or a peak blood alcohol concentra-

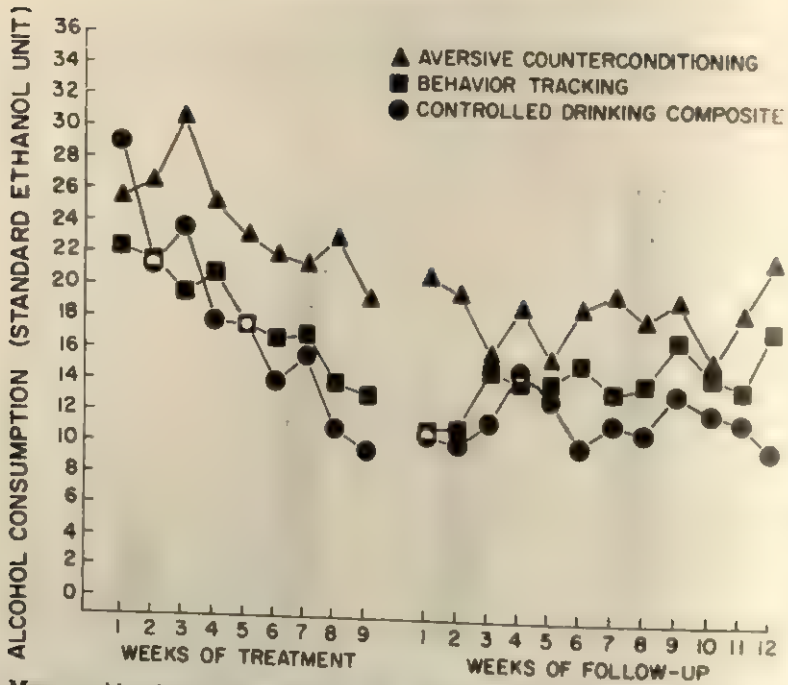


Figure 3. Mean weekly alcohol consumption from daily record cards during treatment and follow-up. (From Chapter 6 "Behavioral Self-control Training in the Treatment of Problem Drinkers" by William R. Miller. In *Behavioral Self-management*, Richard B. Stuart (Ed.). Copyright by Brunner/Mazel (1977). Reprinted by permission.)

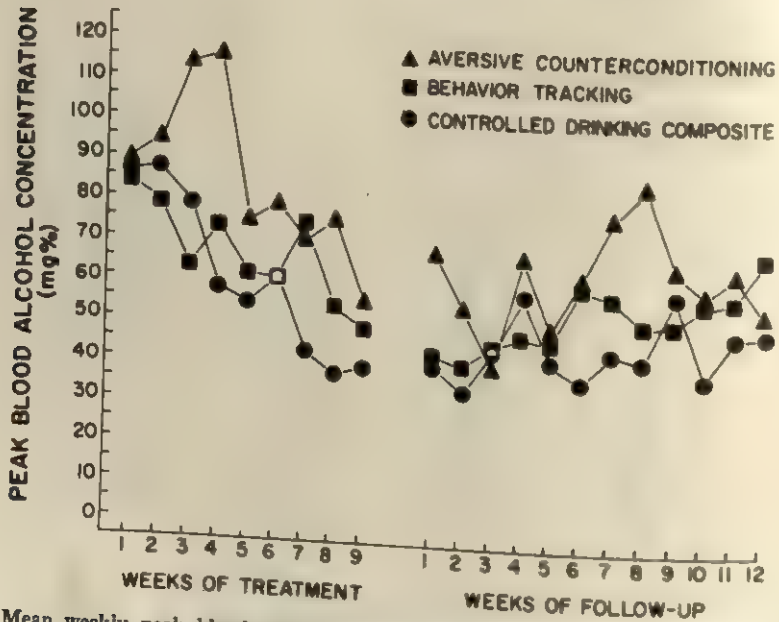


Figure 4. Mean weekly peak blood alcohol concentration from daily record cards during treatment and follow-up. (From Chapter 6 "Behavioral Self-control Training in the Treatment of Problem Drinkers" by William R. Miller. In *Behavioral Self-management*, Richard B. Stuart (Ed.). Copyright by Brunner/Mazel (1977). Reprinted by permission.)

tion of 70 mg% or both, as reported by any data source. Those who, by all reports, were not exceeding these limits were classified as "controlled" at intake and were then assigned to one of two ratings. Controlled clients were rated as *improved* if they showed a noncontradicted reduction of at least 30% in alcohol consumption. ("Noncontradicted" indicates that self-report, reports of significant others, and daily record cards were all in agreement.) Clients who showed less than a 30% reduction or whose reports were contradicted by any data source were rated as showing *no change*.

Problem drinkers were assigned to one of four improvement ratings. *Considerably improved* were those clients who showed at least a 50% noncontradicted reduction in alcohol consumption or who met the controlled drinking criteria following treatment. (Abstainers fell into this category, but they are indicated separately as *abstinent*.) *Moderately improved* clients were those who showed at least a 30% noncontradicted reduction or a decrease greater than 50% that was contradicted by another data source. Clients were rated as *slightly improved* if they showed at least a 10% noncontradicted decrease or a 30%-50% reduction that was contradicted. Clients

with less than a 10% reduction or a contradicted reduction of less than 30% were rated as *not improved*.

Table 2 summarizes the number and percentage of clients assigned to each improvement rating at termination and follow-ups.

Drinker classification. Still another short-hand method for evaluating individual outcome is to determine how many reached an absolute criterion of controlled drinking. The above criteria of not exceeding 20 standard ethanol control units or 70 mg% per week defined the controlled clients, with the additional constraint that these reports could not be contradicted. Number of clients classified as problem, controlled, and nondrinkers at intake, termination, and follow-ups is reported in Table 3.

Drinking Behavior of "Control" Groups

For comparative purposes clients who were excluded from or dropped out of treatment were also evaluated. Although much less information was available for these clients and since they proved more difficult to reach for follow-up, sufficient data were obtained to assign drinker classifications at 3- and 12-month follow-ups. These are reported in Table

Table 2
Number and Percentage of Clients Assigned to Each Improvement Rating at Termination and Follow-ups

Improvement rating	At termination			At 3-month follow-up			At 1-year follow-up		
	AC	BT	CD	AC	BT	CD	AC	BT	CD
Problem drinkers at intake									
Abstinent	0	0	2 (15)	0	1 (7)	2 (15)	2 (20)	1 (7)	1 (8)
Considerably improved	4 (40)	8 (57)	8 (62)	3 (30)	9 (64)	4 (31)	3 (30)	7 (50)	5 (38)
Moderately improved	3 (30)	3 (21)	3 (23)	2 (20)	3 (21)	3 (23)	2 (20)	3 (21)	1 (8)
Slightly improved	0	2 (14)	0	1 (10)	0	0	0	1 (7)	1 (8)
Not improved	3 (30)	1 (7)	0	4 (40)	1 (7)	2 (15)	1 (10)	0	1 (8)
Insufficient data	0	0	0	0	0	2 (15)	2 (20)	2 (14)	4 (31)
Controlled drinkers at intake									
Improved	2	2	1	1	1	1	0	0	1
No change	2	1	1	3	2	1	4	3	1

Note. Numbers in parentheses are percentages. AC = aversive counterconditioning; BT = behavioral self-control training; CD = controlled drinking composite.
From Chapter 6 "Behavioral Self-control Training in the Treatment of Problem Drinkers" by William R. Miller. In *Behavioral Self-management*, Richard B. Stuart (Ed.). Copyright by Brunner/Mazel (1977). Reprinted by permission.

Table 3
Number of Clients in Each Treatment Group Assigned to Each Drinker Classification at Intake, Termination, and Follow-ups

Drinker classification	Assessment period											
	Intake				Termination				3 months			
	AC	BT	CD		AC	BT	CD		AC	BT	CD	
Abstinent	0	0	0		0	0	2 (13)		0	1 (6)	2 (14)	1 (6)
Controlled drinker	4 (29)	3 (18)	2 (13)		6 (43)	11 (65)	10 (67)		4 (29)	11 (65)	7 (50)	9 (53)
Problem drinker	10 (71)	14 (82)	13 (87)		8 (57)	6 (35)	3 (20)		10 (71)	5 (29)	2 (14)	3 (18)
Insufficient data	0	0	0		0	0	0		0	0	3 (21)	4 (24)

Note. Numbers in parentheses are percentages. AC = aversive counterconditioning; BT = behavioral self-control training; CD = controlled drinking composite.

4. Most notable is the high proportion of abstainers among those excluded (and thus treated by the ATSC). Correlations between self-report and significant others' reports were moderately high at both 3-month, $r(29) = .775$, $p < .001$, and 12-month, $r(37) = .480$, $p < .005$, follow-ups.

Examination of the driving records of court-referred clients at the 12-month follow-up revealed that fewer than 10% had repeated their offense of driving while intoxicated. Type of treatment received (ATSC vs. behavior therapy) had no significant effect on level of recidivism.

Effects of Treatment on Other (Nondrinking) Measures

Clients treated by the ATSC and those receiving behavior therapies showed substantial improvement on measures of personal adjustment. On the Profile of Mood States, all three behaviorally treated groups improved significantly ($p < .01$) on all six subscales over the course of treatment and follow-up. ATSC clients showed similar improvement through the 3-month follow-up, but by the 12-month follow-up they had relapsed to pretreatment levels on all subscales. Self-referred dropouts, the only untreated group, were found to be unimproved at all follow-up interviews. Behaviorally treated clients showed significant decreases on 8 of 10 clinical subscales of the MMPI, and similar improvement was found among all control groups.

Effects of a Manual on Maintenance of Controlled Drinking

Clients were randomly chosen at termination to receive or not receive a manual designed to improve maintenance of treatment gains (Miller & Muñoz, Note 2). Of the 26 who received it, 17 indicated at the 3-month follow-up that they had read all or parts of it. Alcohol consumption (from record cards) of clients who read, did not read, or did not receive the manual is shown in Figure 5.

Two-way analyses of variance were used to compare clients who read the manual with those who did not receive it. At termination,

Table 4
 Drinker Classification Within "Control" Groups at 3-Month and 12-Month Follow-ups

Classification	Court-referred						Self-referred dropouts	
	Excluded		Dropouts		Random controls			
	3 months	12 months	3 months	12 months	3 months	12 months	3 months	12 months
Abstinent	23 (66)	17 (49)	0	0	1 (13)	0	2 (13)	2 (13)
Controlled drinker	7 (20)	10 (29)	4 (36)	4 (36)	4 (50)	4 (50)	0	3 (20)
Problem drinker	1 (3)	1 (3)	3 (27)	3 (18)	0	1 (13)	7 (47)	6 (40)
Insufficient data	4 (11)	7 (20)	4 (36)	5 (45)	3 (38)	3 (38)	6 (40)	4 (27)

Note. Numbers in parentheses are percentages.

when manuals were distributed, these groups did not differ on alcohol consumption. At the 3-month follow-up, however, significant differences were found in both weekly consumption, $F(1, 16) = 4.40$, $p < .05$, and weekly peak blood alcohol concentration, $F(1, 16) = 6.21$, $p < .05$. Repeated measures analyses of variance also showed large main effects of the manual ($p < .001$) on both measures. A somewhat surprising aspect of these data is the low level of alcohol use among clients who received but did not read the manual. It appears that these were clients who had already achieved a considerable degree of control by termination, and so they "didn't bother" to read it.

Discussion

Treatment Outcome

The pattern of outcome across treatment modalities is quite consistent. All three treatment groups showed significant reduction in alcohol use. In general the AC group was least improved on all measures throughout the 3-month follow-up, but by 12 months these clients were showing gains comparable to those of the BT and CD groups. A possible explanation of this lies in the self-control manual, given to half of all clients only after the 3-month follow-up. For AC clients this manual contained new methods and informa-

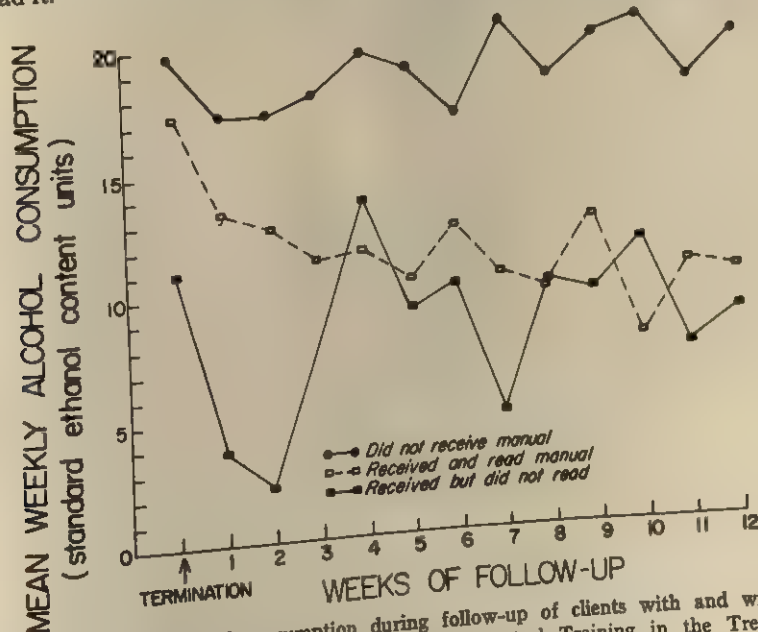


Figure 5. Mean weekly alcohol consumption during follow-up of clients with and without the self-control manual. (From Chapter 6 "Behavioral Self-control Training in the Treatment of Problem Drinkers" by William R. Miller. In *Behavioral Self-management*, Richard B. Stuart (Ed.). Copyright by Brunner/Mazel (1977). Reprinted by permission.)

tion, whereas BT and CD clients had covered this material during treatment. Subsequent research (Miller, Gribskov, & Mortell, Note 3) has indicated that clients using this manual even without therapist contact can significantly reduce their drinking. The seeming reversal among AC clients between 3 and 12 months may thus be partly attributable to use of the manual.

Although no absolute differences between BT and CD were found, the pragmatics of these two treatments would argue for the use of BT, given equal effectiveness. The CD program required the use of expensive equipment: a bar, alcohol, and electric shock. BT requires none of these, is conducted in a standard therapy room, and requires less than 25% of the therapist time needed for CD. BT is also amenable to a wider variety of delivery modes including group therapy, classroom and media presentation, and self-instruction (Miller & Muñoz, 1976). In short, BT is a considerably more cost-effective and flexible approach.

The lack of difference between BT and CD modes is particularly noteworthy because the latter included all of the components of the former. CD clients received instruction in self-monitoring, rate reduction, and functional analysis of drinking behavior—the formal components of BT. In addition CD clients received discriminated aversive counterconditioning and specialized rate control training with in-session drinking and avoidance learning. These additional components apparently made little difference.

These findings underline an important point—that at least in the area of problem drinking, more extensive therapeutic programs are not necessarily more effective. It is often assumed that the addition of new components will increase the overall effectiveness of a program. This issue is a timely one with the present growing emphasis on “multimodal” and “broad spectrum” approaches (Hamburg, 1975; W. R. Miller, 1976). The utility of adjunctive procedures should be evaluated from a cost-effectiveness viewpoint (which may or may not correspond to statistical significance). Expensive additional procedures that add little to a basic treatment program

can then be discarded, and more economical and/or effective ones can be adopted.

Court-referred clients receiving behavior therapies were, because of screening procedures, less severe problem drinkers at intake than were self-referred clients. Nevertheless, court clients responded to treatment in a manner similar to that of their heavier-drinking self-referred counterparts. Both groups reduced their drinking during treatment and maintained gains during follow-up. In no case did treatment mode interact with referral source. For these reasons all clients were combined for analyses of treatment outcome.

A treatment adjunct evaluated within the present study is the self-control manual designed to improve maintenance (Miller & Muñoz, Note 2). Clients who received and read the manual showed better maintenance than did those not given the manual. This effect was a small though significant one, but the use of a self-control manual is a very economical procedure and may thus be justifiable for even small increments in improvement and maintenance (Christensen, Miller, & Muñoz, 1978).

Comparison with Control Groups

Although the present study did not include untreated controls, comparisons are possible with clients who dropped out or received alternative therapies. Of the 15 self-referred clients who refused or dropped out of treatment, only 5 were found to be abstinent or controlled at the 1-year follow-up. This group also showed no improvement on measures of personal adjustment. Clients treated by ATSC, in contrast, showed marked improvement on both drinking and psychological measures. The high frequency of abstinence among ATSC clients suggests that treatment outcome may be related to the ideology and expectations of the treating agency.

Regarding the absolute effectiveness of controlled drinking therapies, judgment must be reserved pending further research. Studies to date have not provided comparable *untreated* controls. It is possible, however, to compare the results of the present study with those of previous outcome research. Emrick (1974)

surveyed the outcome literature on alcoholism and found that the average percentage of clients becoming controlled drinkers was 5.8%, whereas an average of 33.8% became abstainers following treatment. This is the approximate pattern shown by ATSC clients, although the clinic's outcome was well above average in both categories. The pattern of outcome from behaviorally treated groups, however, differs substantially from Emrick's norms, particularly in the controlled category. The overall improvement rate of 84% at termination exceeded Emrick's norms by one standard deviation.

Finally, it is noteworthy that the present study supports the longstanding finding that outcome at a 3-month follow-up is reasonably predictive of the longer range picture with regard to drinking and personal adjustment. Of the 23 "successful" (considerably or moderately improved) cases at the 3-month follow-up for whom 12-month data were also available, 21 (91%) retained their successful status. Four of the six unsuccessful cases at 3 months were rated as successful at 1 year. With regard to drinker classification, 17 of 20 controlled drinkers at 3 months remained so at 1 year, whereas 12 of 17 problem drinkers became controlled. Of the 3 abstainers at 3 months, 1 remained so, 1 was drinking moderately, and 1 had relapsed to heavy drinking.

Conclusions

Certainly behavioral self-control training deserves further study as a treatment method for problem drinkers. A relatively economical procedure oriented toward control rather than abstinence, it offers considerable promise as an early intervention method and may also prove beneficial in the primary and secondary prevention of problem drinking (Muñoz, 1976).

The role of aversive counterconditioning in the control of drinking is more questionable. Aversion therapy, even accompanied by self-monitoring, produced the slowest and smallest initial gains of the methods studied. A counterconditioning component apparently contributed little to our multimodal (CD)

program. These findings suggest that aversive procedures can be discarded without substantially reducing effectiveness (cf. Caddy & Lovibond, 1976; Hamburg, 1975).

The present study also demonstrates the feasibility of paraprofessionals as therapeutic agents for problem drinkers. Previously inexperienced and nonalcoholic therapists effectively trained clients in self-control, producing changes comparable or superior to those of most alcohol treatment outcome studies. Apparently neither an advanced degree nor a history of alcoholism is required for effectiveness, at least with these particular treatment methods.

Further research is needed to determine the optimal combination of treatment components for the control of overdrinking. The present study supports the effectiveness of training clients in self-monitoring, rate control, functional analysis, and blood alcohol concentration discrimination, though the relative contribution of these elements is unknown. Other promising procedures include training in principles of self-reinforcement, basic education regarding alcohol and its effects, involvement of spouse or other family members in treatment, and training in alternatives to the use of alcohol (e.g., systematic desensitization, assertion training). The utility of these additional components remains to be demonstrated.

Reference Notes

1. Miller, W. R. *Behavioral approaches to the treatment of problem drinking: A manual for paraprofessional therapists*. Unpublished manuscript, University of Oregon, 1974.
2. Miller, W. R., & Muñoz, R. F. *Control over drinking: A self-help manual*. Unpublished manuscript, University of Oregon, 1975.
3. Miller, W. R., Gribskov, C., & Mortell, R. *The effectiveness of a self-control manual for problem drinkers with and without therapist contact*. Unpublished manuscript, University of Oregon, 1976.

References

- Alcoholics Anonymous. *Alcoholics Anonymous: The story of how many thousands of men and women have recovered from alcoholism*. New York: Author, 1955.
- Armor, D. J., Polich, J. M., & Stambul, H. B. *Alcoholism and treatment*. Santa Monica, Calif.: Rand Corporation, 1976.

- Briddell, D. W., & Nathan, P. E. Behavior assessment and modification with alcoholics: Current status and future trends. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification*. New York: Academic Press, 1975.
- Caddy, G. R., & Lovibond, S. H. Self-regulation and discriminated aversive conditioning in the modification of alcoholics' drinking behavior. *Behavior Therapy*, 1976, 7, 223-230.
- Christensen, A., Miller, W. R., & Muñoz, R. F. Paraprofessionals, partners, peers, paraphernalia, and print. *Professional Psychology*, 1978, in press.
- Emrick, C. D. A review of psychologically oriented treatment of alcoholism: I. The use and interrelationships of outcome criteria and drinking behavior following treatment. *Quarterly Journal of Studies on Alcohol*, 1974, 35, 523-549.
- Glass, G. V., Willson, V. L., & Gottman, J. M. *Design and analysis of time-series experiments*. Boulder, Colo.: University of Colorado, Laboratory of Experimental Research, 1972.
- Hamburg, S. Behavior therapy in alcoholism: A critical review of broad-spectrum approaches. *Journal of Studies on Alcohol*, 1975, 36, 69-87.
- Hathaway, S. R., & McKinley, J. C. *The Minnesota Multiphasic Personality Inventory*. New York: Psychological Corporation, 1943.
- Huber, H., Karlin, R., & Nathan, P. E. Blood alcohol level discrimination by nonalcoholics: The role of internal and external cues. *Journal of Studies on Alcohol*, 1976, 37, 27-39.
- Jellinek, E. M. *The disease concept of alcoholism*. Highland Park, N.J.: Hillhouse Press, 1960.
- Lloyd, R. W., Jr., & Salzberg, H. C. Controlled social drinking: An alternative to abstinence as a treatment goal for some alcohol abusers. *Psychological Bulletin*, 1975, 82, 815-842.
- Lovibond, S. H., & Caddy, G. Discriminated aversive control in the moderation of alcoholics' drinking behavior. *Behavior Therapy*, 1970, 1, 437-444.
- Marlatt, G. A. The Drinking Profile: A questionnaire for the behavioral assessment of alcoholism. In E. J. Mash & L. G. Terdal (Eds.), *Behavior therapy assessment: Diagnosis, design, and evaluation*. New York: Springer, 1976.
- McNair, D. M., Lorr, M., & Droppelman, L. F. *Manual for the Profile of Mood States*. San Diego, Calif.: Educational and Industrial Testing Service, 1971.
- Miller, P. M. *Behavioral treatment of alcoholism*. New York: Pergamon Press, 1976.
- Miller, W. R. Controlled drinking therapies: A review. In W. R. Miller & R. F. Muñoz, *How to control your drinking*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.
- Miller, W. R., & Caddy, G. R. Abstinence and controlled drinking in the treatment of problem drinkers. *Journal of Studies on Alcohol*, 1977, 38, 986-1003.
- Miller, W. R., & Muñoz, R. F. *How to control your drinking*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.
- Muñoz, R. F. The prevention of problem drinking. In W. R. Miller & R. F. Muñoz, *How to control your drinking*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.
- Nathan, P. E. Alcoholism. In H. Leitenberg (Ed.), *Handbook of behavior modification and behavior therapy*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.
- Nathan, P. E. An overview of behavioral treatment approaches. In G. A. Marlatt & P. E. Nathan (Eds.), *Behavioral approaches to the assessment and treatment of alcoholism*. New Brunswick, N.J.: Rutgers University Center of Alcohol Studies, 1977.
- Rich, C. C., & Davis, H. G. Concurrent validity of the MMPI alcoholism scales. *Journal of Clinical Psychology*, 1969, 25, 425-426.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Selzer, M. L. The Michigan Alcoholism Screening Test: The quest for a new diagnostic instrument. *American Journal of Psychiatry*, 1971, 127, 1653-1658.
- Sobell, M. B., & Sobell, L. C. Alcoholics treated by individualized behavior therapy: One year treatment outcome. *Behaviour Research and Therapy*, 1973, 11, 599-618. (a)
- Sobell, M. B., & Sobell, L. C. Individualized behavior therapy for alcoholics. *Behavior Therapy*, 1973, 4, 49-72. (b)
- Sobell, M. B., & Sobell, L. C. Alternatives to abstinence: Time to acknowledge reality. *Addictions*, 1974, 21, 2-29.
- Sobell, M. B., & Sobell, L. C. Second year treatment outcome of alcoholics treated by individualized behavior therapy: Results. *Behaviour Research and Therapy*, 1976, 14, 195-215.

Received November 15, 1976 ■

Comparison of Usual and Experimental Patients in a Psychiatric Day Center

Marsha Vannicelli and Stephen Washburn
McLean Hospital, Belmont, Massachusetts

Betty-Jane Scheff
Concord Mental Health Center
Concord, Massachusetts

Richard Longabaugh
Butler Hospital
Providence, Rhode Island

In the course of a previously reported study of inpatient and day hospitalization, 59 seriously ill female psychiatric patients were randomly assigned to an inpatient or a day hospital setting. The present study compares the 29 seriously ill patients randomly assigned to the day hospital with a control group of 34 "usual" day patients. The experimental group showed significantly more improvement from baseline to subsequent time periods in three distinct areas: global mental status, subjective distress, and family adjustment. The controls, on the other hand, spent fewer nights in the hospital, used the hospital facilities significantly less during the first 3 months, and incurred a significantly lower cost for the same period. Two measures—number of social work contacts and amount of time spent in the treatment milieu—indicate that experimentals initially required more staff effort than controls, but at later time periods the reverse was true.

A number of controlled studies, including our own (Washburn, Vannicelli, Longabaugh, & Scheff, 1976), of full versus partial hospitalization have demonstrated that patients randomized to a day hospital setting do as well as or better than those who are placed in an inpatient setting. Our study at McLean Hospital of 93 patients, 59 of whom were randomly assigned to either an inpatient or day hospital setting, indicated that on five measures the day patients fared better than those

who were placed in the inpatient program, and on the remaining nine, they did just as well as the inpatients. These findings support the feasibility of treating a large number of patients, formerly treated as inpatients, in a partial hospital setting at considerably decreased cost. What is not known is how the sicker patients, randomized to a day setting, fare compared to the population of patients normally placed there. Equally important is the emotional and material toll on the families and the partial hospital staff in treating these "unusual" day patients. Is there a trade-off somewhere or hidden cost involved in expanding day services to a population of patients who would normally be treated as inpatients? Are hospital and family resources stretched thin by placing this expanded population in the day hospital? Does this sicker population require greater amounts of community support, more psychotherapy, more extensive use of drugs, greater frequency of ancillary service visits, or more intense milieu interactions?

The purpose of this study was to try to

Research was funded in part by U.S. Public Health Service Grants RO1 MH 17464-01 and MH 17464-03A1 and by McLean Hospital Biological Research Support Grant 5 SO1 RR05484. Some additional financial assistance was received from the Fuller Foundation, Inc., Boston, Massachusetts, and the gift of Ronald and Mary Moir, for which we are grateful. We also appreciate the continuing consultation of Robert Rosenthal of Harvard University and the administrative and secretarial assistance of Ann Montague.

Requests for reprints should be sent to Marsha Vannicelli, Director of Outpatient Clinic, Appleton Treatment Center, McLean Hospital, Belmont, Massachusetts 02178.

answer some of these questions. Specifically, we were interested in (a) the relative outcome of sicker patients randomized to a partial hospital setting compared to those who would normally be placed there and (b) process differences between the two partial hospital groups in terms of (a) impact on the family, (b) involvement of the social worker, (c) quality of patient interaction with treatment milieu, and (d) ancillary therapeutic intervention.

Method

Subjects

Description of subjects, setting, recruitment procedures, and outcome measures have been detailed elsewhere (Washburn et al., 1976) and will only be briefly summarized here. Subjects were 93 middle-to upper-middle-class female patients between the ages of 16 and 72 ($M = 32.9$), admitted to McLean Hospital with a primary diagnosis of functional disorder. Of these, 59 were in the randomized group (30 inpatient controls, 29 day center experimentals) and 34 were in the day center control group.

Initial comparison of the two day center groups¹ revealed differences on two demographic and three baseline pathology variables: Proportionately more families of experimental patients (12/29) than controls (5/34) lived farther than 30 miles from the hospital, $\chi^2(1) = 4.35$, $p < .05$; the experimental sample was significantly younger ($M = 28.7$) than the control group ($M = 42.15$; $t = 4.46$, $p < .001$) and was sicker at baseline on three measures of pathology (see Table 1). The two groups did not differ with respect to religion, socioeconomic status, or number of previous hospital admissions.

Subject Recruitment and Procedure

Experimental day subjects were patients who, following an inpatient workup, would have been assigned to an inpatient setting, but they were not so suicidal, homicidal, or incapable of forming a treatment alliance as to make inpatient care an *absolute* necessity (i.e., they were patients who, for the purposes of this study, were considered to be *possibly* treatable in the day setting).

Control day subjects were (a) patients from the community who applied for admission directly to the day center and (b) patients recommended for the day center from the inpatient units. All such patients were judged to be (a) able to forego acting upon suicidal and destructive impulses, (b) able to travel to the day center, and (c) more likely to benefit from treatment in the day center than in any other setting.

Setting

McLean Hospital is a private, nonprofit psychiatric hospital. Its modern day hospital facility offers program options individualized to meet specific patient needs including individual, group, drug and activity therapies, family counseling, and milieu meetings. Program staff includes psychiatrists, social workers, psychiatric nurses, and mental health workers. Twenty-five to 40 patients attend the center daily, with a staff/patient ratio of about 1 to 4. Day center patients who, for brief periods, have difficulty managing in the community at night are housed on an unlocked night care unit.

Instruments and Outcome

Fourteen outcome measures were derived by clustering procedures: (a) Psychiatric Status Schedule—Subject Form (PSS—SF) global mental status, (b) subjective distress, (c) impulse control (Measures a–c from the PSS—SF; Spitzer, Endicott, & Cohen, 1966a), (d) Psychiatric Status Schedule—Informant Form (PSS—IF) global mental status (from PSS—IF; Spitzer, Endicott, & Cohen, 1966b), (e) Psychiatric Evaluation Form (PEF) global mental status, (f) role functioning (Measures e–f from the PEF; Spitzer, Endicott, Mesnikoff, & Cohen, 1966), (g) intrapsychic functioning (adapted from the Camarillo Dynamic Assessment Scales; May & Dixon, 1969), (h) family adjustment—subject, (i) family adjustment—informant, (j) community adjustment (Measures h–j based on Meltzoff and Blumental's, 1966, Outpatient Adjustment Rating scales), (k) burden evaluation (a project-developed graphic rating scale indicating the extent to which the patient's illness imposes a burden on the family), (l) attempted roles (informant's assessment of the number of roles the patient has tried to fill during each 6-month interval), (m) direct charges, and (n) days of attachment (calculated at 3-month intervals, the latter by summing the total days billed for at least a quarter day use of the hospital program).

Instruments and Processes

Informant Forms I and II. Based on the work of Freeman and Simmons (1963), Sainsbury and Grad (1968), and Meltzoff and Blumenthal (1966), these questionnaires were completed by the informant (a family member or friend involved in the day-to-day life of the patient) at 6-month intervals.

Informant Form I assessed the patient's effect on the family as reflected in three summary variables: (a) project assessment of burden, (b) informant subjective burden, and (c) burden imposed by the patient's core pathology.²

¹ Data from the 30 inpatient controls have been previously presented (Washburn et al., 1976).

² Details regarding composite items in these scales are available on request.

Informant Form II focused on the informant's ability to utilize hospital and community support systems, as reflected by (a) willingness to contact the hospital, (b) number of staff contacted, (c) intensity of contact (i.e., amount of contact with the most frequently contacted staff member), (d) number of people contacted outside the hospital, (e) breadth of support system (sum of variables b and d), and (f) number of services used.

The *Social Work Form*, completed monthly, provided the social worker's view of the family as reflected in (a) number of contacts about the patient, (b) necessity for social worker to initiate contacts, (c) family ability to support treatment, and (d) family tolerance for deviant behavior.

Activity logs. Each patient maintained an hour-by-hour activity log that was collected weekly by the research team. Using a random numbers table, a 1-week-long log was selected for each patient per month and averaged over the 6-month periods to obtain data on amount of time (hours) (a) in treatment milieu, (b) on hospital grounds, (c) at work and/or school, (d) in psychotherapy, and (e) in other therapies.

Social interaction. Quantity and quality of social exchange was coded at one meeting per week and was averaged over the 6-month periods to determine number of (a) healthy acts, (b) acts directed toward the patient from other patients, (c) acts toward the patient from staff, (d) acts by the patient toward other patients, (e) acts by the patient to staff, and (f) total acts by the patient.

Staff assessment of burden. During the last 9 months of the study, the day center staff rated all patients on a scale of 1 (no burden) to 7 (extreme burden). An average rank was computed monthly for each patient and was then averaged over the 9-month period.

Administration of Measures

Baseline measures of mental status and of family and community functioning were obtained prior to official acceptance into the day center. Initial measures on all other instruments were obtained within a few days after acceptance. Subsequent assessment occurred for up to 2 years.

Results

Outcome Analyses

Fourteen two-way analyses of variance with repeated measures were computed on the outcome variables examining the data over time from our two randomized groups and the day center control group. Planned-comparison *t* tests for both main effects and interactions were computed using the mean square error from the corresponding analysis of variance as the best estimate of error variance.

Main Effects

The planned comparisons for main effects were used to determine whether patients (across treatment groups) tend to get better or worse if examined at baseline and then again at later time periods. The *t* tests showed that *in general, patients improved from baseline to later periods*. These data have been reported in detail (along with the interaction effects comparing the two randomized groups) in Washburn et al. (1976).

Interaction Effects: Experimental Group Compared to the Controls

Table 1 presents the means and *t* values for measures that showed differences between patients randomized to the day center (experimental) compared with patients who normally would be treated in that setting (controls). Two-tailed tests of significance were used to compare the amount of change occurring in the experimental group between two given points in time with the amount of change occurring in the control group during the same time period.³ In addition, absolute time comparisons were made on the two groups at baseline, at 1 year, and for all analyses regarding days of attachment and direct charges.

On the *PEF global mental status*, the experimental group showed significantly more improvement from baseline to the overall posttest period and from baseline to posttest at 3, 6, and 8 months than did the controls. The control group started with significantly ($p < .01$) less psychopathology (baseline $M = 10.82$) than the experimental group (baseline $M = 14.29$) and improved very little over time, whereas the experimental group improved dramatically, becoming healthier than the controls (though not significantly so) at

³ Specific comparisons were as follows: Amount of change between baseline and Posttest 1, Posttest 2, mean of Posttests 1 and 2, the overall postrandomization period (mean of all posttests), and amount of change between Posttests 1 and 2 and Posttests 2 and 3. Where no baseline data were available, parallel comparisons were made with Posttest 1 instead of baseline.

Table 1

Significant Comparisons for Day Center Experimentals Versus Day Center Controls

Outcome measure	Experimental	Control		df
Significant difference on mean change				
Global mental status				
PEF; change from baseline to				
Overall post (through 24 mo.)	4.30 ^a	.15	2.93**	323
3rd post (Mo. 5-6)	3.47 ^a	.58	3.17**	323
6th post (Mo. 12-14)	6.31 ^a	.46	3.23**	323
8th post (Mo. 18-24)	5.24 ^a	.70	2.93**	323
PSS-SF; change from baseline to				
Overall post (through 18 mo.)	8.93 ^a	.96	2.52 ^a	136
Overall first year	9.23 ^a	2.06	2.40 ^a	136
Second 6 months	10.09	1.04	2.88**	136
Subjective distress; change from baseline to				
Overall post (through 18 mo.)	8.97 ^a	.34	2.29 ^a	136
Overall first year	9.28 ^a	1.66	2.14 ^a	136
First 6 months	10.96 ^a	.23	2.87**	136
Significant differences on absolute means				
Days of attachment (1st 3 mo.)	60.59	50.74 ^a	2.12 ^a	282
Direct charges (1st 3 mo.)	\$5,480	\$3,370 ^a	3.27***	60
Global mental status				
PSS-SF (baseline)	48.10	43.27 ^b	2.31 ^a	136
PEF (baseline)	14.29	10.82 ^b	3.70**	323
Subjective distress (1 year)	38.30 ^a	46.38	2.89**	136
Impulse control (baseline)	50.33	46.68 ^b	2.87**	136

Note. PEF = Patient Evaluation Form; PSS-SF = Psychiatric Status Schedule—Subject Form.
^a Superior outcome. ^b Less pathology.
^{*} $p < .05$. ^{**} $p < .01$. ^{***} $p < .001$.

about the 5th month. This pattern of greater improvement in the experimental group is also reflected in the data from PSS global mental status and subjective distress.

An even more pervasive pattern that appears in the PEF and is reflected in nearly all other measures is that of greater initial pathology (baseline) for the experimental group but less pathology at 1 year. In fact, on 10 out of 11⁴ measures, the experimentals appeared sicker than the controls at baseline (and significantly so on PEF impulse control and PSS-SF; see Table 1). In contrast, at 1 year the experimentals appeared better in absolute than the controls on 8 out of 11 measures (significantly so on subjective distress).

Only in terms of *number of days of attachment* and *direct costs* did the experimental group fare worse than the controls. During the first 3 months, the experimentals were attached to the hospital for significantly more days than the controls, and they incurred

⁴ The one exception appears to be a function of the peculiarities of the instrument used. The PSS-IF format is such that pathology must actually be observed to be recorded by the interviewer. Thus, for controls, who were generally living with the informant, any existing pathology is observed and recorded, whereas for experimentals, who were in the hospital and more removed from the informant, the amount of observed pathology is generally less. The Informant Form of the PSS is not an appropriate measure under these circumstances (a position with which Spitzer and Endicott now agree).

significantly higher costs. At no other time period, however, did either the cost or the number of days of attachment differ.

Process Analyses

Using the variables described under Process Measures, 27 two-way analyses of variance with repeated measures were computed to examine the data over time (baseline, 6 months, 1 year, and 18 months) for our three groups.⁵

Whereas a planned-comparison *t*-test strategy was used for the outcome analyses, the process variables were analyzed using a more traditional analysis of variance strategy, with *t*-test comparisons made on absolute scores only when *F* ratios were significant.

Main Effects: Group Differences Between Experimentals and Controls

Nights in the hospital was the only process measure that revealed significant main effects differences, $F(1, 60) = 10.03$, $p < .01$, between the two day center groups—The usual patients spent fewer nights ($M = 7.05$) in the hospital across time than the experimentals ($M = 21.98$). (The interaction effects discussed below are helpful in understanding this main effect.)

Interaction Effects: Experimental and Control Groups Compared over Time

Significant interaction effects differentiating control from experimental patients at specific time periods emerged on 8 of the 27 variables analyzed.

The meaningful⁶ significant comparisons are summarized below for each variable.

Nights in the hospital. Both experimental and control patients dropped significantly in number of nights spent in the hospital after the first 3 months ($F = 47.76$, $p < .01$). The control patients, however (who initially, and at all subsequent time periods, spent significantly fewer nights than experimentals), dropped off much less during later periods.

Intensity of contacts within the hospital. At 6 months⁷ informants involved with experimental patients had significantly more

contacts with one primary care agent than informants associated with controls, $F(6, 129) = 3.44$, $p < .01$.

Number of outside people contacted. At 18 months informants involved with experimental patients contacted significantly fewer people than informants involved with controls, $F(6, 125) = 2.78$, $p < .05$.

Social work contacts about the patient. At baseline the social worker made significantly more contacts about experimental patients than about controls, whereas at 1 year the reverse was true, $F(6, 97) = 3.35$, $p < .01$. Over time the control group did not change in terms of amount of contacts the social worker made, whereas the experimental group showed a downward trend with significant drops in number of social work contacts from baseline to each subsequent time period, and also between 6 months and each subsequent period.

Necessity for the social worker to initiate contact. The social worker felt a greater need to initiate contacts about the experimental patients than about controls at both 1 year and 18 months, $F(6, 89) = 4.33$, $p < .001$.

Both *social work contact about the patient* and *the necessity to initiate contact* show that more effort was initially expended by the social worker for experimental patients than for controls. With regard to the number of contacts a reversal later occurred—the social worker at 1 year had more contacts about the controls than about the experimentals.

Family ability to support treatment. At 18 months (but not before), families of control patients were seen by the social worker

⁵ On two measures, nights in the hospital and time spent in treatment milieu, data were examined over only three time periods (no baseline data) and for only the two day center groups.

⁶ Significant differences between the two day groups at two different times, for example, are not discussed unless in the context of other data they add meaning. Presentation of means, statistical procedures, and rationale of procedures used are detailed in a more complete manuscript, available on request.

⁷ For all analyses that follow, 6-month data reflect the amount of activity between baseline and 6 months; 12-month data reflect activity between 7 and 12 months, 18-month data reflect activity between 13 and 18 months.

as having more ability to support treatment than were families of experimentals, $F(6, 83) = 6.79, p < .01$.

Time spent in treatment milieu. Both experimental and control groups spent significantly more time in the treatment milieu during the first 6 months than at subsequent periods, $F(2, 122) = 7.19, p < .01$. Of greater interest, whereas experimentals initially (Months 1-6) spent more time in the milieu than controls, during both subsequent 6-month periods controls spent significantly more time than did experimentals.

Drug and staff burden analyses. Chi-squares examining the proportion of drug users and nonusers at 6 months and 12 months for each major drug category (major tranquilizers, minor tranquilizers, antidepressants, hypnotics, and sedatives) and t tests on staff burden ratings revealed no differences between experimentals and controls.

Discussion

Our outcome comparison of usual day patients with sicker patients experimentally treated in the day setting showed a significantly greater decrease in pathology for the experimental patients. At first glance the greater change found in experimental patients may seem less than impressive, since these sicker patients had more room to improve. These findings are more striking, however, when it is noted that on 8 out of 11 measures, the experimentals not only improved more, but at 1 year they were also healthier *in absolute* than the controls. Only in terms of direct hospital charges and number of days hospitalized did the experimentals fare worse than the controls.

These findings suggest that a sicker range of patients may be more responsive to the day setting than those normally placed there. The experimentals exhausted their outpatient treatment providers to the extent that inpatient evaluation was required. There is likely to be a lively, even turbulent, interaction with these patients that counters the apathy seen in day settings for the chronically ill.

Our process data confirm this—More effort was expended by the social worker, who, particularly at baseline, was having (and initiating) more contact about experimental patients

than about controls. At 18 months the social worker still felt more pressure to initiate contact about the experimentals although actually having fewer contacts about them than about controls. At that time the experimental families were also contacting significantly fewer people outside the hospital than were control families. By then the families of experimentals were apparently feeling less need for support (perhaps due to the dramatic improvement of their sicker member), but the social worker remained vigilant.

The initial greater exertion by the social worker for experimental patients can perhaps be best understood in terms of the flamboyant illness of these patients in the early phases of hospitalization and their greater propensity to change. This propensity to change may also contribute to the social worker's feeling of greater need to initiate contact about experimental patients and her assessment that their families were less able to support treatment. She was apparently feeling the pressure of change to a greater extent than were the families themselves.

It is important to note that experimentals not only did as well or better regarding pathology than controls—but that this occurred at no greater overall cost. It is true that experimentals, initially, are harder to handle (as indicated by hospital charges, nights spent in the hospital, days of attachment, amount of time in the treatment milieu, and pressure for social work). But after this initial period, a reversal occurred with regard to the two latter measures, with controls later requiring more attention than experimentals. In addition, in terms of 20 other important variables reflecting, (a) impact on the family, (b) interaction with the treatment milieu, and (c) need for ancillary therapeutic intervention (drugs, psychotherapy, etc.), there were no significant differences between the two groups.

The greater initial pressure imposed on the social worker by the experimental patients, greater number of days attached to the hospital, and corresponding higher charges during the first 3 months can be partially explained in terms of our study design. At the time of assignment to the day center, the experimental group had spent up to 6 weeks as inpatients and were caught in the inertia

of the hospital system. In addition, community support systems had been disrupted and needed to be remobilized—a process complicated by the fact that families of experimental patients tended to live further from the hospital than those of controls. Extensive social work effort was often needed to establish alternative living arrangements before these patients could leave the hospital.

In contrast, the community support network of the controls had been only minimally disrupted. Thus, there was less initial chaos, lower pressure on the social worker, fewer nights of hospitalization and lower hospital charges for the control group, and ultimately *less pressure toward change or improvement*. As systems theory would predict, a system is more open to change when change is already occurring; and the more acute state of the experimental patient requires immediate change. Although positive change counters chaos, it also causes further change (creating in its wake new chaos and need for additional change). The control patient, on the other hand, is part of a more static system in which fewer changes are demanded, creating less chaos and, in turn, less pressure for change (and improvement).

Other less comprehensive explanations of the experimental group's superior outcome are possible but less persuasive. The gains of experimental patients may be related to some extent to the specific effects of their initial inpatient status and their greater subsequent use of night care. Alternatively, it may be that because of their younger age and/or a possible Hawthorne effect, the experimentals received more attention than the controls and responded to this greater attention with improved performance. The social work data are consistent with the latter explanation. However, nursing staff in the day hospital, when specifically questioned, were not able to identify which patients were experimentals and which were controls, nor were there any differences in terms of staff interventions with experimental and control patients. Thus, if a Hawthorne effect were operative, it would have been an indirect function of differences in the behavior of the social worker (who worked with the significant others, rather than with the patient)—the social worker's atten-

tion affecting family members, who, as a consequence, acted differently to facilitate improvement in the patient. This explanation begins to overlap with the systems theory that we have proposed.

Perhaps the best alternative explanation for the observed differences between the two day hospital groups is that control patients were more chronically ill than experimentals. Although there was no difference between the two groups on one index of chronicity (number of prior hospitalizations), the control patients were significantly older. Their age increases the possibility that they may have been more chronically ill—perhaps due, at least in part, to years of insufficient systems demand for change.

Only further research can sort out the extent to which our results can be accounted for by these overlapping explanations. However, whatever the explanation, our data suggest that expansion of day services to include patients who have normally been treated as inpatients may require more initial effort, but this effort pays off.

References

- Freeman, H. E., & Simmons, O. G. *The mental patient comes home*. New York: Wiley, 1963.
- May, P. R., & Dixon, J. The Camarillo Dynamic Assessment Scales. *Bulletin of the Menninger Clinic*, 1969, 33, 1-35.
- Meltzoff, J., & Blumenthal, R. L. *The day treatment center: Principles, application, and evaluation*. Springfield, Ill.: Charles C Thomas, 1966.
- Sainsbury, P., & Grad, J. The effects that patients have on their families in a community care and a control psychiatric service—A two-year follow-up. *British Journal of Psychiatry*, 1968, 114, 265-278.
- Spitzer, R. L., Endicott, J., & Cohen, G. *Psychiatric Status Schedule, Subject Form*. New York: Biometrics Research, New York State Department of Mental Hygiene, 1966. (a)
- Spitzer, R. L., Endicott, J., & Cohen, G. *Psychiatric Status Schedule, Informant Form*. New York: Biometrics Research, New York State Department of Mental Hygiene, 1966. (b)
- Spitzer, R. L., Endicott, J., Mesnikoff, M., & Cohen, G. M. *Psychiatric Evaluation Form, Form R2*. New York: Biometrics Research, New York State Department of Mental Hygiene, 1966.
- Washburn, S., Vannicelli, M., Longabaugh, R., & Scheff, B-J. A controlled comparison of psychiatric day treatment and inpatient hospitalization. *Journal of Consulting and Clinical Psychology*, 1976, 44, 665-675.

Received January 27, 1977 ■

Comparative Effectiveness of Day Hospital and Inpatient Psychiatric Treatment

W. E. Penk and H. L. Charles

Veterans Administration Hospital, Dallas, Texas

T. A. Van Hoose

Southern Methodist University, Dallas

The question, "Are treatment effects of partial hospitalization comparable to effects of full-time hospitalization?" was studied by means of a pretreatment-posttreatment control group design. Ratings of home and community adjustment were obtained from community informants (relatives or close friends) for 24 day hospital clients (partial hospitalization condition) matched with 24 inpatient clients (full-time hospitalization condition), as well as from an unmatched sample of 79 consecutive inpatient admissions. All groups evidenced improvement 2 months after treatment started, particularly on measures of symptom reduction. Although patterns of improvement were comparable for both settings in univariate data analyses, multiple discriminant function analysis indicated that the day hospital sample evidenced greater gains in attentiveness and in employment. Gains favoring the day hospital sample were more striking for the unmatched three-group comparison than for the matched two-group comparison. The findings indicate that partial hospitalization is an attractive alternative to inpatient psychiatric hospitalization; further, differences in subject characteristics are influential in studies of treatment outcome.

Efforts to evaluate different systems for delivering mental health services are increasing (e.g., Struening & Guttentag, 1975). Program evaluation methodology, however, is still developing: Investigators continue to question which measures reliably and validly assess treatment outcome, which perspectives in treatment outcome are preferred, and whether

assessment should concern general treatment effects or an individual's attainment of specific goals (e.g., Schulberg, Sheldon, & Baker, 1969).

Despite uncertainties, program evaluation based on multiple measures obtained from another person significant to the client in treatment has emerged as a reliable and valid source of data (e.g., Ellsworth, 1975). Systematic observations of home and community adjustment, rated by a person close to the client or so-called "community informant," seem to provide an acceptable and important perspective for inferences about treatment effects.

The purpose of the present study was to compare, from the community informant's perspective, differential effectiveness of two approaches in psychiatric treatment: day hospital (or part-time) treatment versus inpatient (or full-time) psychiatric treatment. Although day hospital and inpatient treatment approaches share in common many program-

This research was supported by the Medical Research Service of the Veterans Administration. Preliminary findings were reported at the annual meeting of the American Psychological Association, Honolulu, Hawaii, September 1972. The authors are indebted to the former Hospital Director, E. P. Whitaker, Veterans Administration Hospital, Dallas, Texas, who gave assistance in data collection, and to Isham Kimbell, Chief, Psychiatry Service, Veterans Administration Hospital, Dallas, Texas, who provided diagnostic and psychiatric ratings for inpatient psychiatry admissions.

Requests for reprints should be sent to W. E. Penk, Veterans Administration Hospital, Dallas, Texas 75216. Penk is also at the Psychology Division, University of Texas Health Science Center, Dallas, Texas.

matic features (such as individual and group therapy), the two approaches differ in amount of time a client spends at the respective treatment locale—8 hours a day for the day hospital setting versus 24 hours a day for the inpatient psychiatric setting. Supposedly, less time spent in the day hospital disrupts the client's home and community involvement less, whereas more time spent in an inpatient unit, so the notion goes, fosters overdependency on the institution and interferes with home and community social ties (e.g., Jones, 1953).

Clinical impressions have supported the view that day hospitalization is an effective treatment alternative to inpatient hospitalization (e.g., Chasin, 1967). Exploratory experimental efforts have indicated that favorable treatment outcome is associated with the part-time system of mental health services (e.g., Guy, Gross, Hogarty, & Dennis, 1969); specifically, day hospitalization seems to disrupt the client's employment status less (cf. Herz, Endicott, Spitzer, & Mesnikoff, 1971). Such findings were observed, however, in studies flawed by methodological shortcomings, such as absence of control groups or comparisons of samples differing in subject characteristics notably influential of treatment outcome. As program evaluation methods have improved, investigators have become less certain that differences in treatment systems contribute to differences in treatment outcome (e.g., Ellsworth, Note 1; Ellsworth, Finnell, & Leuthold, Note 2).

The present study compared day hospital and inpatient differential treatment effects under conditions in which subject characteristics were controlled. The experimental strategy comprised both a "fixed scale" and a "fixed time" approach—fixed scale in the sense that all subjects were compared for outcome differences on the same set of multiple measures of psychosocial functioning derived from ratings by community informants assessing clients' home and community adjustment before and after treatment, and fixed time in the sense that pretreatment and posttreatment time duration did not vary from subject to subject.

Method

Experimental Strategy

A modified form of the pretreatment-posttreatment control group design served as the experimental strategy (cf. Campbell, 1969, Design 4). Day hospital clients comprised the experimental group; inpatient clients, matched for relevant demographic characteristics with day hospital clients, served as the primary control group. The matched inpatient sample was selected from 22% of total inpatient admissions during the 2-month time period of the study (in 1972). A second control group sample was drawn from the remaining number (78%) of consecutive admissions. This second "unmatched" inpatient control group consisted of those inpatients not matched with the day hospital subjects but admitted to the inpatient unit during the time of the study. The unmatched inpatient sample furnished an additional perspective for analyzing sample comparability of the matched day hospital-inpatient cohorts, particularly whether constraints had been introduced by matching in sampling on the "external generalizability" (cf. Campbell, 1969) of the findings.

Subject Selection

Matching of day hospital and inpatient samples followed a cohortlike pairing procedure. A board-certified psychiatrist, attending intake staffing conferences in both settings, paired those inpatient admissions who seemed similar to day hospital admissions in age, type, and severity of personal problems and comparable resources in coping with conflicts. The process of subject pairing seemed free of subjective and institutional biases, particularly since there were no fixed criteria for day hospital admission at the time data were collected. Although the samples appear to be representative of clients in inpatient and outpatient psychiatric units in general medical and surgical Veterans Administration hospitals, several categories of subjects (or 9% of admissions during the study period) did not participate, that is, the physically infirm with psychological disorders, those with severe chronic brain disorders, and those who were acutely suicidal or homicidal.

Additional steps were followed to insure comparability of matched day hospital and inpatient samples on background characteristics and current levels of intelligence and personality functioning. Each of the matched subjects was rated by his attending psychiatrist on the Brief Psychiatric Rating Scale (BPRS; Overall & Gorham, 1962); samples did not differ significantly on the 18 scales. Matched subjects did not differ on measures obtained from a brief psychological screening battery (which included a biographical inventory, the Shipley Institute of Living Scale, Shipley, 1940, and several standardized measures of personality variables). Matched subjects did not differ on informants' reports about

home and community resources and background (e.g., income, living arrangements, treatment outcome expectancy, etc.).¹

The matched samples averaged 32 years of age, completed approximately 11½ years of education, and scored within the average range of intelligence (Wechsler Adult Intelligence Scale equivalency scores = 108). Nearly two thirds of both samples were married; half were employed. Forty-eight percent of both samples were diagnosed as schizophrenic, and average length of program participation was about 33 days.

Program Evaluation Instrument

The Personal Adjustment and Role Skills scale (PARS, Ellsworth, 1975) served as the measure of treatment outcome. The PARS consists of 57 items and is scored for eight factor-analytically derived measures of home and community adjustment. Reliability and validity of PARS scales are well established; manuals are available, and scale definition and development have been detailed elsewhere (Ellsworth, 1975; Ellsworth, Foster, Childers, Arthur, & Kroeker, 1968). Seven scales were analyzed in the present study: (a) Interpersonal Interaction, which registers reactions to social interaction; (b) Agitation-Depression, which measures feelings of pessimism; (c) Anxiety, which assesses tenseness; (d) Confusion, which evaluates attentiveness and efficiency in mental concentration; (e) Alcohol and/or nonprescribed drug use, which estimates moderation in alcohol or drug use; (f) Social Activity, which reflects the number of social contacts (e.g., attending movies or visiting friends); and (g) Employment, which gauges amount and participation in productive, gainful employment. An eighth scale, Parent-Child Interaction, was omitted from analysis, since it did not apply to all subjects. The first five scales are described as measures of symptoms; Social Activity and Employment are regarded as measures of instrumental role activity.

Procedure

Veterans were asked shortly after admission to participate in a study evaluating effectiveness of their treatment. Each man who volunteered (indicated by signing a consent form after reading details of the research plan and the PARS) nominated a relative (or close friend) who could be contacted by mail for information about home and community adjustment before treatment ("pretreatment" phase) and then again 2 months after treatment started (posttreatment phase). Thirty-seven pairs of matched day hospital and matched inpatient admissions agreed to participate; 35 day hospital and 37 inpatient informants returned scorable pretreatment scales (pre-PARS), and 24 day hospital and 24 inpatient informants returned posttreatment scales (post-PARS) 2 months later. The matched samples, then, consisted of 24 day hospital-inpatient pairs for whom

both pre- and post-PARS, BPRS, and group psychological screening battery scores were available. (No significant differences were found for pre-PARS scores when comparing the group returning pre-PARS only with the group returning both pre- and post-PARS.)

Informants were sought also for the unmatched inpatient sample. One hundred thirty inpatient admissions (or 95% of the unmatched inpatient sample) agreed to participate: 108 informants returned scorable pre-PARS and 79 mailed back both pre- and post-PARS scales. These 84% pre-PARS and 62% post-PARS return rates were higher than the expected rate of return (cf. Ellsworth, 1975; Fontana & Dowds, 1975).

Treatment Settings

Programmatic activity was similar in both treatment settings. The main difference was time spent in the unit—8 hours a day for the day hospital versus 24 hours a day for the inpatient unit.²

The general medical and surgical hospital that contains both units is located in a large Southwestern metropolitan complex. The day hospital was situated in a single-purpose unit located on the hospital grounds but separate from the main complex. A daily structured program of individual and group therapy was offered along with regularly scheduled periods for occupational and recreational therapy. Staff ratio was 1:4; five staff members (psychiatrist, nurse, social worker, psychologist, and secretary) had an average daily patient load of 20.

The inpatient unit, housed in the main general medical and surgical complex, consisted of four wards and four treatment teams. Staff ratio was 1:4, with five staff positions (psychiatrist, nurse, nursing aide, with halftime social worker, psychologist, occupational and recreational therapists) responsible for approximately 20 patients. The inpatient treatment program consisted of scheduled activities similar to those found in the day hospital, including ward self-government meetings.

Research Questions

Three questions were addressed from the stance of the pretreatment-posttreatment control group design. First, Do post-PARS ratings differ from pre-PARS ratings for the matched day hospital and inpatient samples combined? In other words, Was there evidence of improvement following 2 months of

¹ Supplementary tables are available containing descriptive and inferential statistics for psychiatrists' ratings, psychological test scores, and informants' ratings of home information.

² A supplementary table describing type and amount of time spent in treatment activity, diagnosis, and length of stay is available from the first author.

treatment? Second, does the pattern of improvement (or pre- and post-PARS differences) vary between the matched day hospital and inpatient samples? In other words, Does treatment outcome differ as a function of treatment setting? Third, Do findings for the matched day hospital and inpatient samples hold when compared with pre- and post-PARS differences for the unmatched inpatient sample? In other words, Do the findings generalize?

Results

Patterns of Improvement for Matched Groups Combined

Matched day hospital and inpatient samples combined evidenced significant gains in correlated-means *t*-test comparisons of four of the seven PARS scores. For the measures of symptoms but not instrumental role skills (Table 1), increased Calm-contentedness, $t(47) = 5.44, p < .001$; for increased Attentiveness, $t(47) = 2.18, p < .03$; for lessened Anxiety, $t(47) = 6.76, p < .001$; and for moderation in Alcohol Abuse, $t(47) = 4.37, p < .001$.

Treatment Effects for Separate Programs

Day hospital. The day hospital sample, considered separately by correlated-means *t* tests (Table 2), evidenced increased Calm-contentedness, $t(23) = 4.51, p < .001$; increased Attentiveness, $t(23) = 3.03, p < .001$; lessened Anxiety, $t(23) = 6.18, p < .001$; and increased moderation in Alcohol Abuse, $t(23) = 2.02, p < .05$.

Matched inpatient sample. The matched inpatient sample, comparing pretreatment and posttreatment PARS scale differences by correlated-means *t* tests, evidenced similar changes: increased Calm-contentedness, $t(23) = 3.30, p < .003$; lessened Anxiety, $t(23) = 3.65, p < .001$; and increased moderation in Alcohol Abuse, $t(23) = 4.22, p < .001$. The inpatient sample gain for Attentiveness, unlike the day hospital scores, did not reach an acceptable level of significance ($\alpha = .05$).

Unmatched inpatient sample. The unmatched inpatient sample evidenced less magnitude in change than either of the two matched groups; moreover, the unmatched inpatient group sustained a significant loss in

Table 1
Pre- and Post-PARS Normative Scores for Day Hospital and Inpatient Samples Combined

Variable	Pre-PARS		Post-PARS	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Interpersonal involvement	47.77	8.41	50.17	8.20
Calm-contentedness	39.50	10.42	47.29	11.27
Attentiveness	38.54	12.25	43.06	11.56
Anxiety	38.31	7.94	47.52	9.11
Alcohol abuse	42.42	13.21	49.15	9.42
Social activity	43.06	7.95	42.42	7.76
Employment	29.35	15.68	29.92	16.28

Note. PARS = Personal Adjustment and Role Skills scale. $n = 48$ for both pre-PARS and post-PARS.

Employment (Table 2). Correlated-means *t*-test differences were significant for increased Interpersonal Interaction, $t(78) = 2.26, p < .03$; increased Calm-contentedness, $t(78) = 5.22, p < .001$; increased Attentiveness, $t(78) = 3.51, p < .001$; lessened Anxiety, $t(78) = 5.17, p < .001$; moderation in Alcohol Abuse, $t(78) = 3.99, p < .001$; and decline in Employment, $t(78) = 2.78, p < .007$.

Differential Comparisons of Matched Samples

The matched day hospital and inpatient samples were comparable on six of seven pre-PARS scales, differing only in greater moderation in Alcohol Abuse for the day hospital sample, independent-means $t(46) = 2.38, p < .03$. Post-PARS independent-means *t*-test comparisons, however, resulted in a larger separation between the two matched groups. The day hospital sample was higher in Calm-contentedness, $t(46) = 1.78, p < .08$; Social Activity, $t(46) = 2.93, p < .005$; and Employment, $t(46) = 2.57, p < .03$. The two matched groups did not differ significantly in gain score comparisons (i.e., when comparing separately the difference between each pre- and post-PARS scale score).³

³ Simple difference (or gain) scores were analyzed following recommendations of Fontana and Dowds (1975).

Table 2

PARS Normative Scores for Matched Inpatient and Day Hospital and Unmatched Inpatient Samples

Variable and time	Day hospital ^a		Matched inpatient ^a		Unmatched inpatient ^b	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Interpersonal interaction						
Pre	48.54	6.43	48.21	8.60	47.95	8.21
Post	49.96	8.66	50.38	7.88	49.44	9.05
Gain ^c	1.42	7.72	2.17	8.60	1.49	4.51
Calm-contentedness						
Pre	42.04	8.94	36.96	11.34	41.19	11.68
Post	50.13	9.92	44.46	12.01	46.56	11.48
Gain	8.09	8.78	7.50	11.34	5.37	9.92
Attentiveness						
Pre	37.17	8.58	37.42	12.21	39.86	11.48
Post	44.04	11.86	42.08	11.42	43.68	10.78
Gain	6.87	11.11	4.66	12.20	3.82	10.07
Anxiety						
Pre	38.46	9.04	38.17	6.86	38.48	9.77
Post	49.42	9.62	45.63	8.33	43.76	11.34
Gain	10.96	8.69	6.46	6.86	5.28	9.39
Alcohol abuse						
Pre	46.42	11.50	38.48	13.83	45.27	12.50
Post	50.46	9.31	47.83	9.55	50.62	10.33
Gain	4.04	9.89	9.41	13.83	5.35	12.99
Social activity						
Pre	45.25	8.31	40.88	7.09	42.68	8.29
Post	46.46	7.45	40.38	6.95	43.32	8.90
Gain	1.41	8.98	-0.50	7.09	0.64	9.22
Employment						
Pre	30.63	16.59	28.08	14.96	27.82	15.43
Post	35.63	17.15	24.21	13.42	23.44	14.25
Gain	5.00	18.43	-3.87	14.96	-3.38	13.69

Note. PARS = Personal Adjustment and Role Skills scale.

^a *n* = 24.

^b *n* = 79.

^c Gain scores include both negative and positive numbers; accordingly, standard deviation values may be larger than mean values in some instances.

Both groups achieved significant gains. Improvement, however, does not vary widely as a function of differences in health delivery system. Three findings in the univariate statistical analysis suggest greater, albeit mild, advantages for the day hospital experience: (a) the significant increase in Attentiveness for the day hospital sample when differences were analyzed separately for each setting, (b) greater gains in all PARS scores considered collectively but not separately, and (c) higher post-PARS Social Activity and Employment for the day hospital sample after

nonsignificant pre-PARS (except for Alcohol Abuse). Such trends, discernible by univariate analysis, reinforced a need for multivariate analysis. Further analysis was justified a posteriori by finding significant intercorrelations among the PARS scales.⁴ Multiple discriminant analyses (Klecka, 1975) were performed separately for pre-PARS, post-PARS, and gain scores (a) in two-group comparisons be-

⁴ A supplementary table of pre- and post-PARS scale intercorrelations is available on request.

tween matched day hospital and inpatient samples and (b) in a three-group comparison among matched inpatient and day hospital samples and the unmatched inpatient group.

Multivariate Analysis of the Two Matched Samples

The matched samples did not differ for either pre-PARS or gain scores, whereas significant differences were found for post-PARS scores. The day hospital sample scored higher on Social Activity, univariate $F(1, 46) = 8.56, p < .001$, and for Employment, $F(1, 46) = 6.60, p < .001$. One discriminant function significantly separates the two matched groups, eigenvalue = .612; canonical correlation = .616; Wilk's $\Lambda = .620$; $\chi^2(6) = 21.01, p < .002$. Less depression (reverse scoring of Calm-contentedness) and more Social Activity and Employment differentiates the day hospital sample from the inpatient group. (See Table 3 for standardized discriminant coefficients and group centroids.)

Multivariate Analysis of the Three Matched and Unmatched Samples

Results indicate a more favorable outcome for the day hospital sample when multivariate analysis was performed in a three-group comparison, that is, matched inpatient and day hospital with unmatched inpatient sample. First, the three groups did not differ in univariate and multivariate comparisons on pre-

PARS scores. Second, significant differences occurred in both univariate and multivariate post-PARS scale comparisons; the day hospital sample obtained significantly higher post-PARS Employment scores, $F(2, 127) = 6.58, p < .001$, and Social Activity scores, $F(2, 127) = 3.21, p < .05$. Multivariate analysis yielded two significant discriminant functions (see Table 3 for coefficients and group centroids): The first obtained an eigenvalue of .218; canonical correlation = .423; Wilk's $\Lambda = .749$; $\chi^2(6) = 11.47, p < .05$. The first discriminant function varied along a continuum from disruption in thinking (interferences in Attentiveness) to Employment and lessened Anxiety; the second function varied from depression and social inactivity to lessened Anxiety. Post-PARS group differences were characterized by rated increases in Employment and Attentiveness (with which the day hospital sample aligned) and decreases in Anxiety (with which the inpatient sample aligned).

The three groups differed significantly on PARS gain scores. The day hospital sample evidenced a greater decrease in Anxiety, univariate $F(2, 127) = 3.29, p < .05$, and a greater increase in Employment, $F(2, 127) = 3.48, p < .05$. The discriminant function for gain scores, although nonsignificant ($p < .06$), was comparable to the first discriminant function for post-PARS scores; that is, the day hospital sample was higher in posttreatment Employment and lower in posttreatment Anxiety.

Table 3
Discriminant Function Coefficients

Variable	Two-group comparison coefficient	Three-group comparison	
		First coefficient	Second coefficient
Interpersonal involvement	.431	-.051	.674
Calm-contentedness	-.862	1.43	-.882
Attentiveness	.834	-.972	-.181
Anxiety (lessened)	.084	.722	1.001
Alcohol abuse (moderation)	-.252	-.091	-.553
Social activity	-.760	.334	-.614
Employment	-.787	.882	-.034
Group centroid			
Matched day hospital	-.766	.958	-.101
Matched inpatient	.766	-.068	-.643
Unmatched inpatient		-.261	-.158

Discussion

Informants regarded clients in both settings as improving in home and community adjustment. Such improvement was registered more in measures of symptom reduction than in measures of instrumental role skills (cf. Ellsworth, 1975; Fontana & Dowds, 1975).

Generality of the findings was limited by the comparatively narrow pretreatment and posttreatment time interval. Previous research suggests, however, that most variance is accounted for by patients' level of adjustment 1 month after discharge (Ellsworth & Schoonover, Note 3). The present findings thus may reflect a persisting effect of treatment outcome.

The present strategy of using two control groups in comparing differential treatment effectiveness indicated that the matching procedure did not isolate a sample that was unrepresentative of those seeking treatment; the results may be extended to the treatment setting in general. Other questions were raised, however, by this research tactic. Why were group differences more pronounced for the unmatched group comparison than for the "matched" group comparison? Obviously, outcome gains were maximized in the three-group comparison but minimized in the two-group comparison. Although it was not possible to identify factors associated with this phenomenon, the observation that degree of improvement varies as a function of differences in background characteristics of samples demonstrated the importance of describing samples in detail (cf. Campbell, 1969).⁵

Although treatment gains were achieved by both the day hospital and the inpatient samples, the evidence suggested, nevertheless, a more favorable outcome under the day hospital condition. The trend favoring the day hospital experience was barely discernible in the univariate analysis but was delineated in the multivariate analysis by the emerging pattern of gains in employment and social activity unique for day hospital clients. The result was not unanimous, for gain score differences (barely) failed to show a similar trend. Finding significantly less disruption in employment for the day hospital sample is consonant

with findings in earlier studies (e.g., Herz et al., 1971; Ellsworth et al., Note 2).

If one accepts a more prudent interpretation—that day hospital treatment is at least an acceptable alternative, but not necessarily a better alternative, to inpatient treatment for most clients seeking inpatient or outpatient treatment—then the question about which modality is the treatment of choice could be debated from the standpoint of other considerations. One such consideration is treatment "costs."

The day hospital approach recommends itself further based on its comparatively lower cost. Direct costs for each day hospital client averaged \$33.26 per day (or \$39.02 for indirect costs) compared with \$43.16 a day for each inpatient client (or \$73.49 including indirect costs).⁶ Cost parameters such as salaries, housing, equipment, and so on, do not mean necessarily that the "costs" of treatment have been estimated in their entirety, however. "Cost" of treatment, like the cost of living, is a multidimensional construct whose amount is determined not only by the assessment of material outlay but also by weighing the extent to which family, friends, and client are taxed in the process of coping with emotional distress. The use of community informants when evaluating outcome taps a significant dimension in determining cost. The evidence supports the notion that at least from the perspective of the community informant, day hospitalization is an important treatment alternative for many people.

⁵ Diagnosis (e.g., schizophrenic vs. nonschizophrenic) did not emerge as a significant outcome variable (cf. Erickson, 1976).

⁶ The authors wish to thank John Molnar-Subajda, Management Analyst, Veterans Administration Hospital, Dallas, Texas, for computing average daily costs for day hospital and inpatient psychiatry units.

Reference Notes

1. Ellsworth, R. B. *Initial and outcome differences between community and clinic clients and psychiatric patients*. Paper presented at the annual meeting of the American Psychological Association, Chicago, August 1975.
2. Ellsworth, R. B., Finnell, K. C., & Leuthold, C. *The comparative effectiveness of community, intensive, and traditional programs in treating psy-*

chiatric patients. Manuscript submitted for publication, 1976.

3. Ellsworth, R. B., & Schoonover, R. *Aftercare services in psychiatric treatment*. Manuscript submitted for publication, 1976.

References

- Campbell, D. T. Factors relevant to the validity of experiments in social settings. In H. C. Schulberg, A. Sheldon, & F. Baker (Eds.), *Program evaluation in the health fields*. New York: Behavioral Publications, 1969.
- Chasin, R. M. Special clinical problems in day hospitalization. *American Journal of Psychiatry*, 1967, 123, 779-785.
- Ellsworth, R. B. Consumer feedback in measuring the effectiveness of mental health programs. In M. Guttentag & E. L. Struening (Eds.), *Handbook of evaluation research* (Vol. 2). Beverly Hills, Calif.: Sage Publications, 1975.
- Ellsworth, R. B., Foster, L., Childers, B., Arthur, G., & Kroeker, D. Hospital and community adjustment as perceived by psychiatric patients, their families, and staff. *Journal of Consulting and Clinical Psychology Monograph*, 1968, 32(5, Pt. 2).
- Erickson, R. C. Length of stay and adjustment and role skill changes of day hospital patients. *Newsletter for Research in Mental Health and Behavioral Sciences*, 1976, 28, 15-18.
- Fontana, A. F., & Dowds, B. N. Assessing treatment outcome. *Journal of Nervous and Mental Disease*, 1975, 161, 221-230.
- Guy, W., Gross, M., Hogarty, G. E., & Dennis, H. A controlled evaluation of day hospital effectiveness. *Archives of General Psychiatry*, 1969, 20, 329-338.
- Herz, M. I., Endicott, J., Spitzer, R. L., & Mendikoff, A. Day versus inpatient hospitalization: A controlled study. *American Journal of Psychiatry*, 1971, 127, 1371-1382.
- Jones, M. *The therapeutic community*. New York: Basic Books, 1953.
- Klecka, W. R. Discriminant analysis. In N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, & D. H. Bent (Eds.), *Statistical package for the social sciences*. New York: McGraw-Hill, 1975.
- Overall, J. E., & Gorham, D. R. The brief psychiatric rating scale. *Psychological Reports*, 1962, 10, 799-812.
- Schulberg, H. C., Sheldon, A., & Baker, F. (Eds.) *Program evaluation in the health fields*. New York: Behavioral Publications, 1969.
- Shipley, W. C. A self-administering scale for measuring intellectual impairment and deterioration. *Journal of Psychology*, 1940, 9, 371-377.
- Struening, E. L., & Guttentag, M. (Eds.) *Handbook of evaluation research* (Vol. 1). Beverly Hills, Calif.: Sage Publications, 1975.

Received March 3, 1977 ■

Test Anxiety and the Passage of Time

Irwin G. Sarason and Rick Stoops
University of Washington

Three experiments were performed dealing with the relationship of test anxiety and achievement-oriented instructions to time perception. After being given either achievement-orienting or neutral instructions, subjects waited for an undesignated period of time and then performed an intellectual task. The dependent measures were subjects' estimates of the duration of the waiting and performance periods and their scores on the assigned task. High-test-anxious subjects' time estimates were significantly greater than the estimates of the other subjects, and their performance was at a relatively low level. Evidence is presented supporting the hypothesis that highly anxious persons under stress experience cognitive interference and preoccupation that makes time pass slowly and results in poor performance. The implications of the findings are discussed particularly in terms of the need for training programs capable of fostering improved cognitive skills requiring self-control.

Although the generalization that time is precious holds for many situations, it does not always seem to have validity. There are situations in which time drags and one wishes it were possible to speed up the clock. For the football team behind 30-0, time moves slowly, whereas for the team on the way to victory, time flies and its players desire more time to win "big." Similar differences in subjective time estimates seem to hold for many types of situations and events, a critical factor being the character of the particular situation. For example, waiting for what may be bad news about a loved one who is in the hospital can be excruciating, and there is evidence that time passes very slowly for depressives (Bech, 1975).

Although the literature on time estimation is sizable, much of the work done has focused on time estimation as a function of either personality characteristics (such as anxiety) or experimental conditions (see Meade, 1966;

Siegmán, 1962). In analyzing subjective judgments of time duration, it seems logical to consider simultaneously two variables: the situation and the characteristics (hopes, fears, etc.) that a person brings to the situation (Buchwald & Blatt, 1974; Sarason, Smith, & Diener, 1975). The present article reports the results of three experiments devised from this perspective. An important feature of the experiments is the inclusion of two types of data—time estimations and performance.

In these experiments subjects were told that they would be administered a test of intelligence but that there would first be a waiting period. (There were also control groups not given the achievement-orienting communication.) How does time pass while awaiting the evaluation? Obviously the way in which the intellectual evaluation is construed has a bearing on the answer to this question. Many psychological instruments reflect aspects of the construing process and the meanings that individuals attach to particular types of situations. Scales dealing with well-defined situations, such as those designed to tap test and speech anxiety, can be viewed as measuring cognitive activity and worrying, which is stimulated by a specific event or demand (such as having to take a test or make a speech).

This research was supported in part by Office of Naval Research Contract N00014-75-C-0905 NR 170-804. We are indebted to John Granach and Hans F. N. Crombag for their assistance.

Requests for reprints should be sent to Irwin G. Sarason, Department of Psychology, University of Washington, Seattle, Washington 98195.

The personality measure used in the three experiments was the Test Anxiety Scale (TAS; Sarason, 1972). High scorers on this measure have been shown to perform more poorly than others on difficult, complex tasks administered under achievement-orienting conditions that emphasize the evaluation of one's performance (Sarason, 1975). Test anxiety can be interpreted as a form of self-preoccupation—characterized by self-awareness, self-doubt, and self-depreciation—that influences overt behavior and psychological reactivity. Other types of anxiety may be similarly interpreted. The self-preoccupying thoughts of the highly anxious individual interfere with adaptation at several points in the course of information processing. They narrow or otherwise influence the attentional focus on environmental cues; distort encoding, transformation, and planning strategies; and influence responses that may be selected to cope with challenges confronting the individual.

Available evidence suggests that the relatively poor performance of highly test-anxious persons under achievement-orienting conditions is not due to low intelligence but rather to the cognitive interference of a personalized, self-centered approach to evaluational situations. The expectations of a highly anxious person seem to be different from those of others (Doris & Sarason, 1955). When this person performs poorly, it may not be due merely to cognitive interference and self-preoccupation during the test. It may also be related to the time spent *anticipating* the test with dread. These personalized anticipations contribute to inefficient, ineffective preparation for the test.

The experiments reported here were aimed at providing information about the way in which persons differing in anxiety fill time. It was predicted that in the presence of achievement-orienting cues, time would pass more slowly for high-anxiety scorers than for middle- and low-anxiety scorers. When these cues are not present, there should not be a significant gap in estimates of time duration among groups differing with regard to test anxiety. Furthermore, the effects of an achievement orientation should be as noticeable while the individual is waiting to perform

as during performance itself. The first two experiments deal with these hypotheses and differed in the length of the waiting period preceding performance. In the third experiment, the performance period was greatly lengthened, and a specially prepared postexperimental questionnaire was administered in an attempt to clarify the relationships among achievement orientation, test anxiety, and cognitive interference. It was expected that highly anxious persons who are underachievers would describe themselves as having more task-irrelevant thoughts than would persons with middle and low anxiety scores.

Experiment 1

Method

Subjects. The subjects were 48 male and 48 female students from introductory psychology classes at the University of Washington. All subjects were approximately 18–19 years of age. Prior to and independent of the experiment, 550 students had been administered the TAS (Sarason, 1972). The subjects were drawn from the top and bottom 15% of the distribution of TAS scores and from a group in the middle of the distribution. The subjects in the high TAS group had scores of 26 and above; subjects in the low TAS group had scores of 9 and below. The middle TAS group had scores between these cutoff points. Subjects' assignments to experimental conditions were random within the requirements of the experimental design. (This method of assignment to conditions was followed in all three experiments.)

Procedure. After being escorted into the experimental room, subjects were asked to put their watches out of sight until the end of the session because, they were told, a watch might be a distracting stimulus during the experiment. Each subject was given the task of writing a short self-description and a description of one other person. They were allotted 3.5 minutes for each description. The subjects, who were run individually, were then given either achievement-orienting or neutral instructions. The achievement-orienting instructions were given as follows:

The test you are about to take is part of a widely used intelligence test. This is the most crucial part of this study, since it will be used to give me a measure of your intelligence. I have to get the materials. I'll be back shortly.

The experimenter then left the room and shut the door. The experimenter returned 2 minutes later with the test materials and asked the subject to write down (in minutes and seconds) the length of time that he or she had been sitting alone. The experimenter then continued with the following:

As I said, the test you are about to take is part of an intelligence test. This test has been found to

predict such things as course grades, success in later life, and, to some extent, the kind of personality you possess. Of course, your own intelligence will primarily determine whether you do well or poorly on the test. At a later date you will have an opportunity to compare your IQ score with those of the other people in this study. You will then be able to determine how your abilities and capacities compare with other people like you.

The subject was then given a difficult version of the digit symbol task (variations of the letter *L*) with the following instructions: "The purpose of this task is to put the symbols in the numbered boxes as prescribed by the code at the top of the page. Try the three examples." The materials were an adaptation of those used by Sarason and Palola (1960). The subject then worked for 3.5 minutes on the digit symbol task and was asked to write down (in minutes and seconds) the length of time that he or she had been working on the test. The subject was then debriefed and excused.

The neutral instructions were as follows: "I have to get the materials we need. I'll be back shortly." The experimenter then left, returned in 2 minutes, and asked the subject to write down (in minutes and seconds) the length of time that the subject thought the experimenter had been gone. The experimenter then gave instructions on how to perform the digit symbol task. The subject worked for 3.5 minutes, after which he or she was asked to write down (in minutes and seconds) the amount of time that he or she had worked on the task. The subject was then debriefed and excused.

Results

A $3 \times 2 \times 2$ analysis of variance (encompassing test anxiety, sex, and conditions) of subjects' time estimates while waiting to do the digit symbol task failed to yield statistically significant results. However, the results for the Test Anxiety \times Conditions interaction were in the expected direction, $F(2, 84) = 3.73$, $.05 < p < .10$. The mean waiting time estimate of the high TAS subjects in the experimental group was 156.8 sec, whereas the mean for the high TAS control group was 127.7 sec. On the other hand, the middle and low TAS control and experimental group means were in the opposite direction. When the time estimates were grouped into intervals defined by overestimates, underestimates, or exact estimates of the actual time period and a chi-square analysis was performed, anxiety and conditions were found to interact significantly, $\chi^2(10) = 19.32$, $p < .05$. High TAS subjects overestimated the interval during which they believed they were waiting to take

a test to a greater extent than those in the high TAS control group. More low and middle TAS subjects overestimated the interval during which they believed they were waiting to perform on a neutral task than did low TAS subjects, who believed that they were waiting to take a test. The results for time estimates of the period in which subjects performed were in the same direction as for the waiting period but were not statistically significant.

An analysis of variance of the digit symbol performance scores yielded one significant result, that for the test anxiety main effect, $F(2, 84) = 4.07$, $p < .05$. The low TAS mean (92.8) was higher than the middle (82.2) and high (81.5) means. There were no significant sex differences in either Experiment 1 or 2.

Experiment 2

Although in several respects the results of Experiment 1 were consistent with expectations, they tended to be weak and in some instances inconsistent, for example, the fact that the middle TAS group's performance scores more closely resembled those of the high TAS group than those of the low TAS group. In the hope of uncovering more decisive relationships, a second, related experiment was performed. Two changes pertained to the temporal variable. Because the 2-minute waiting period in Experiment 1 might not have been long enough to allow for significant effects of the test anxiety and experimental variables to show up, the waiting period in Experiment 2 was lengthened to 4 minutes. In addition, subjects performed the digit symbol task for 4 instead of 3.5 minutes.

Another change in Experiment 2 was the task on which subjects worked prior to performing the digit symbol task. Instead of writing a short self-description and a description of one other person, subjects performed for 7 minutes on an anagrams task. This type of concept-formation task was deemed somewhat more consistent than the writing task, since the experimental emphasis was on the evaluation of intellectual performance. For the achievement-oriented group, the anagrams were so difficult that it was certain that no subject could complete the task in the allotted time. For the control group, the anagrams

were relatively easy, and all subjects successfully completed the task. The changes made in the preliminary task (particularly its difficulty level and time pressure) were designed to heighten stress on the evaluation of performance among subjects in the achievement-orientation group.

Method

Subjects. The subjects were 120 undergraduates at the University of Washington. The 60 males and 60 females were divided into groups on the basis of their scores on the TAS using the same cutoff points as were used in Experiment 1.

Procedure. The experiment followed a $3 \times 2 \times 2$ analysis of variance design. The variables were (a) TAS—high, middle, and low scorers; (b) conditions—achievement-orienting and neutral control; and (c) sex—male and female subjects.

This experiment used the procedures of Experiment 1 except for the following changes: (a) Before performing the digit symbol task, subjects worked on anagrams (easy ones for those in the control group, difficult ones for the experimental or achievement-orientation group), and (b) the waiting period and time for performance on the digit symbol test were 4 minutes each.

Results

The analysis of variance for the subjects' estimates of the waiting period prior to performing on the digit symbol task yielded two significant results—the effects for test anxiety, $F(2, 108) = 3.57, p < .05$, and conditions, $F(1, 108) = 5.03, p < .01$. The test anxiety result reflected larger waiting period estimates for the high (303.8) than for the low (274.1) and middle TAS (269.5) groups. The larger high TAS estimates were mainly attributable to the high TAS group that received the achievement-orienting condition. This is shown by the fact that the mean for this group was 337.6, whereas the high TAS neutral group's mean was 270.0, $F(1, 38) = 4.31, p < .05$. Table 1 presents the mean waiting time estimates together with the mean estimates of time spent on the digit symbol task and performance scores on that task. Because there were no significant sex differences, male and female results were combined in Table 1.

The significant effects in the waiting period analysis were also significant in the analysis of subjects' estimates of time spent on the

Table 1
Mean Waiting Time and Task Time Estimates and Digit Symbol Performance Scores in Experiment 2

Condition	Waiting time ^a	Task time ^a	Performance
H-E	337.6	346.3	68.5
H-C	269.9	261.9	87.8
M-E	279.0	258.1	100.4
M-C	260.0	259.8	98.6
L-E	285.0	266.8	100.6
L-C	253.3	258.5	102.6

Note. H, M, and L refer to high, middle, and low levels of test anxiety, respectively; E and C refer to experimental and control conditions.

^a In seconds.

digit symbol task. The TAS main effect, $F(2, 108) = 5.13, p < .01$, was due to higher estimates for the high (304.1) than for the low (262.7) and middle (258.9) TAS groups. Again, the greater high TAS mean was due mainly to the high TAS achievement-oriented group. The mean for this group was 346.3, whereas the comparable low TAS control group mean was 261.9. The TAS \times Conditions interaction, $F(2, 108) = 7.81, p < .001$, was attributable to differences between the high TAS (346.3) achievement-oriented group and all other groups in the experiment (combined $M = 261.0$).

The analysis of digit symbol performance scores yielded two significant findings: for test anxiety, $F(2, 108) = 7.82, p < .001$, and Test Anxiety \times Conditions, $F(2, 108) = 3.21, p < .05$. The main effect for test anxiety was due to poorer performance for the high than for the middle and low TAS groups. This in turn was explicable largely in terms of the relatively poor performance of the high-scoring group. The high TAS achievement-orientation mean was 68.5; the mean for the high TAS control group was 87.8; and the mean for all middle and low TAS groups combined was 100.5. These results contributed to the significant TAS \times Conditions interaction.

Experiment 3

The procedural changes made in Experiment 2 led to more clear-cut results than were

Table 2
*Mean Waiting Time and Task Time Estimates,
 Anagram Performance Scores, and Cognitive
 Interference Scores in Experiment 3*

Condition	Waiting time ^a	Task time ^a	Ana-grams score	Cognitive inter-ference score ^b
H-E	357.0	1354.1	3.3	33.2
H-C	286.5	1114.0	4.8	24.6
M-E	266.3	1031.5	5.5	18.2
M-C	274.4	1103.5	5.7	21.6
L-E	266.5	1172.0	5.0	19.8
L-C	265.0	1140.5	5.0	21.4

Note. H, M, and L refer to high, middle, and low levels of test anxiety, respectively; E and C refer to experimental and control conditions.

^a In seconds.

^b Reflects the degree to which the subject reported experiencing interfering thoughts.

obtained in Experiment 1. The findings of the two investigations support the conclusions that not only is the performance of high TAS subjects deleteriously affected by achievement-orienting instructions, but, in addition, these subjects tend to overestimate both the duration of the test period and the period during which they wait to have their ability evaluated. This seems analogous to the tendency to exaggerate the time spent in the dentist's waiting room and in his or her office. Anticipating and going through unpleasant, frightening, or threatening experiences seem to take up a lot of time. If this interpretation is correct, the question arises: Do persons differing in anxiety fill time periods in similar or dissimilar ways? Experiment 3 was designed to provide evidence relevant to this question and to extend the generality of results obtained in Experiments 1 and 2.

In Experiment 3 the tasks used in Experiment 2 were reversed. All subjects worked on a digit symbol task prior to a waiting period and then were asked to solve a series of difficult anagrams. The period during which they were occupied with the anagrams was much longer than was the case for the postwaiting task in the earlier experiments. Following performance on the anagrams task, the subjects responded to a questionnaire dealing with their cognitive activity during that task.

Method

Subjects. The subjects were 60 female undergraduates at the University of Washington. They were divided into groups on the basis of their TAS scores, using the same cutoff points that were used in Experiments 1 and 2.

Procedure. The experimental design encompassed two factors: (a) high, middle, and low TAS scores and (b) achievement-orienting and neutral instructions. Each subject worked on the digit symbol task for 4 minutes. This was followed by a 4-minute waiting period. At the end of the waiting period, subjects performed for 18 minutes on a series of difficult anagrams. The experiment concluded with subjects responding to a questionnaire about cognitive activity while occupied with that concept-formation task. The questionnaire was a modified version of one developed by Diener and Endresen (Note 1). It dealt with the tendencies during performance to have task-irrelevant thoughts (e.g., what the experimenter thought about the subject, wondering about how others had done on the task).¹

Results

There were two significant *F*s in the analysis of waiting period time estimates, for test anxiety, $F(2, 54) = 8.31, p < .001$, and for Test Anxiety \times Conditions, $F(2, 54) = 3.31, p < .05$. The high, middle, and low TAS means were 321.8 sec, 270.4 sec, and 266.3 sec, respectively. The interaction result showed that the greater high TAS mean was attributable mostly to the high TAS group receiving achievement-orienting instructions. The mean for that group was 357.0 sec, whereas the high TAS control group mean was 286.5 sec. Table 2 presents the means of the four dependent measures for all groups in Experiment 3.

The analysis of estimates of duration of the anagrams task also yielded two significant *F*s: for test anxiety, $F(2, 54) = 3.29, p < .05$, and for Test Anxiety \times Conditions, $F(2, 54) = 3.41, p < .05$. Again, the significant results were explicable largely in terms of the relatively large estimates given by the high TAS achievement-orientation group (see Table 2). The mean for that group was 1,354.1 sec, whereas the mean for all other groups combined was 1,112.3 sec.

¹ The questionnaire is available from Irwin G. Sarason.

When an analysis was performed on the number of correct responses to the anagrams task, only the test anxiety effect was statistically significant, $F(2, 54) = 3.35, p < .05$. As the means in the third column of Table 2 show, this effect was due mainly to the relatively poor performance of the high TAS group receiving the achievement-orienting instructions.

There were two significant results in the analysis of interfering activity scores that were obtained by summing the subjects' responses to the questionnaire's 11 items. These scores reflect the degree to which the subject reported experiencing interfering thoughts. These significant results were the F for test anxiety, $F(2, 54) = 5.33, p < .01$, and for Test Anxiety \times Conditions, $F(2, 54) = 3.27, p < .05$. As column 4 of Table 2 shows, most of these significant effects were due to the high scores obtained by the high TAS achievement-orientation group, whose mean was 33.2. The mean for the high TAS control group was 24.6, and the combined mean for the middle and low TAS group was 20.3. Results for separate analyses of individual items were in every case in the same direction as the results presented for the questionnaire as a whole.

One item appended to the questionnaire asked the subject to indicate on a 7-point scale the degree to which her mind wandered while working on the anagrams task. An analysis of variance of these scores yielded significant F s for test anxiety, $F(2, 54) = 3.45, p < .05$, and for Test Anxiety \times Conditions, $F(2, 54) = 3.61, p < .05$, the directions of these results resembling those in the other analyses.

General Discussion

Looking at the total picture provided by the findings of the three studies, it appears that persons for whom tests are noxious experiences (high TAS subjects) tend to overestimate, to a greater degree than do others, both the time during which their performance is being evaluated and the period during which they are waiting for the evaluation to take place. Adding to the picture is the fact that high-test-anxious subjects performed at significantly lower levels than did low and middle scorers

when emphasis was placed on the evaluational implications of performance.

The evidence from Experiment 3 concerning cognitive interference is enlightening from the standpoint of what persons think about while working on a task. High test anxious subjects, more so than low and middle scorers, attribute to themselves preoccupations about how poorly they are doing, how other people are faring, and what the examiner will think about the subject. These findings are in line with those obtained by Diener and Endresen (Note 1). It is difficult not to interpret these preoccupations as having the effect of appreciably complicating the task at hand. Although a measure of cognitive interference during the waiting period was not obtained, it seems likely that similar preoccupations would have especially characterized high test anxious subjects at that time.

Janis (1958) has described the "work of worrying" as a step toward dealing effectively with a threatening or challenging reality situation. Arnold (1960) has also referred to worrying as a preparation for action. Although this emphasis on the positive aspects of worry is commendable, sight must not be lost of the important fact of individual differences in worrying. The person who describes himself or herself as *characteristically* being a worrier may not be taking a positive first step in coping with stress when he or she begins to worry. Rather, the individual may be creating subjectively vivid personal fictions and exaggerations that instead of being of help in the coping process, serve to exacerbate or create stress where it otherwise might not exist at all. A high score on a measure of trait anxiety may then be viewed as reflecting obsessive self-preoccupation and thereby the tendency to complicate situations that may already be sufficiently challenging. In the case of the TAS, inferences are drawn only to a defined domain of activity being evaluated.

Doob (1971) has presented a cogent, wide-ranging survey of temporal dimensions of behavior. Further research is needed on the role of a number of temporal variables in stress and anxiety. For example, although high TAS scorers in Experiment 3 described themselves as very much self-preoccupied during the 18-minute-long anagrams task, it may well be

that these covert responses were not evenly distributed throughout that time period. It would be interesting to obtain measures of cognitive interference at several points during performance. Similarly it would be valuable to have a clearer picture of cognitive activity during waiting periods. Breznitz (1971) has called attention to a process of incubation by which the stress value of a stimulus or situation is enhanced during a waiting period. The time interval between warning of an impending threat (e.g., a test) and its actual occurrence merits study as an independent variable. Of equal importance is the variable of the time filler: What happens, if anything, during the waiting interval?

Another problem of both theoretical and practical significance is the matter of how to help people gain more control over their behavior in situations requiring anticipation of, and later coping with, stress. The problem of self-preoccupation and its intrusive effects is not limited to the domain of anxiety. Some self-preoccupied persons worry, others respond covertly and overtly with anger, and still others are suspicious of potential unseen traps in the situations with which they must deal. The rapidly developing fields of cognitive training and cognitive therapy have much to contribute to the analysis, and, where desirable, to a reduction of the tendency to be self-preoccupied (Mahoney, 1974; Meichenbaum, 1972; Rimm & Masters, 1974). Training aimed at strengthening adaptive cognitive skills (e.g., planning a course of action, waiting patiently, and reducing intrusive self-preoccupation) is especially relevant in reactions to personal threat. In challenging situations—either self-imposed, as in climbing a mountain, or unexpected, as in a sudden illness—the utilization of time can be of the utmost importance. Control over one's thoughts may be the decisive factor in successfully meeting a situational challenge.

The results of the series of experiments reported here lend support to the growing interest in a Persons \times Situations approach to personality (Sarason, 1977; Sarason, Smith, & Diener, 1975). Two indices—estimates of the durations of time periods and performance—were found to be a joint function of a situational characteristic: (a) whether or not

emphasis was placed on an achievement orientation and (b) an individual difference variable, test anxiety. Other evidence from diverse fields that supports the need for an interactional psychology is now available (Magnusson & Endler, 1977). To understand and predict behavior, data are needed about both the information provided by environmental situations and the characteristics of persons who must process the information.

Reference Note

1. Diener, E., & Endresen, K. *Task-irrelevant responses of highly test anxious students*. Unpublished manuscript, University of Washington, 1974.

References

- Arnold, M. B. *Emotions and personality* (2 vols.). New York: Columbia University Press, 1960.
- Bech, P. Depression: Influence on time estimation and time experience. *Acta Psychiatrica Scandinavica*, 1975, 51, 42–50.
- Breznitz, S. A. A study of worrying. *British Journal of Social and Clinical Psychology*, 1971, 10, 271–279.
- Buchwald, C., & Blatt, S. J. Personality and the experience of time. *Journal of Consulting and Clinical Psychology*, 1974, 42, 639–644.
- Doob, L. W. *Patterning of time*. New Haven: Yale University Press, 1971.
- Doris, J., & Sarason, S. B. Test anxiety and blame assignment in a failure situation. *Journal of Abnormal and Social Psychology*, 1955, 50, 335–338.
- Janis, I. L. *Psychological stress*. New York: Wiley, 1958.
- Magnusson, D., & Endler, N. S. (Ed.). *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, N.J.: Erlbaum, 1977.
- Mahoney, M. J. *Cognition and behavior modification*. Cambridge, Mass.: Ballinger, 1974.
- Meade, R. D. Achievement motivation, achievement, and psychological time. *Journal of Personality and Social Psychology*, 1966, 4, 577–580.
- Meichenbaum, D. H. Cognitive modification of test anxious college students. *Journal of Consulting and Clinical Psychology*, 1972, 39, 370–380.
- Rimm, D. C., & Masters, J. C. *Behavior therapy: Techniques and empirical findings*. New York: Academic Press, 1974.
- Sarason, I. G. Experimental approaches to test anxiety: Attention and the uses of information. In C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (Vol. 2). New York: Academic Press, 1972.
- Sarason, I. G. Anxiety and self-preoccupation. In I. G. Sarason & C. D. Spielberger (Eds.), *Stress and anxiety* (Vol. 2). New York: Wiley, 1975.

Sarason, I. G. The growth of interactional psychology. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, N.J.: Erlbaum, 1977.

Sarason, I. G., & Palola, E. G. The relationship of test and general anxiety, difficulty of task, and experimental instructions to performance. *Journal of Experimental Psychology*, 1960, 59, 185-191.

Sarason, I. G., Smith, R. E., & Diener, E. Personality research: Components of variance attributable to the person and the situation. *Journal of Personality and Social Psychology*, 1975, 32, 199-204.

Siegmán, A. W. Future-time perspective in the perception of duration. *Perceptual and Motor Skills*, 1962, 15, 609-610.

Received March 15, 1977 ■

A Cognitive-Behavioral Treatment for Impulsivity: A Group Comparison Study

Philip C. Kendall
University of Minnesota

A. J. Finch, Jr.
Virginia Treatment Center for Children,
Richmond

From a clinic population of emotionally disturbed children, 20 children initially identified as impulsive were randomly assigned to either a cognitive-behavioral treatment group or to a control group. The treatment group received six sessions of verbal self-instructions via modeling with response-cost contingent upon errors during training, and controls received similar training without specific treatment. Although two self-report measures and teacher and staff ratings of locus of conflict did not show treatment effects, both an increase in the latency and a decrease in the error measures from the Matching Familiar Figures Test and improved teacher ratings of impulsive classroom behavior revealed positive effects due to treatment. These treatment effects remained evident at follow-up. The present study provides group-comparison evidence for the efficacy of the cognitive-behavioral treatment for modifying impulsivity.

Children are not known for their willingness to consider alternatives when it comes to making a decision. Often, they want the first candy bar they see, the toy nearest them, or anything one of their friends has. Fortunately, most children become more cautious with age. However, some children never seem to stop and think, and reflective reasoning seems alien to them—these are the impulsive children.

The cognitive dimension of reflection-impulsivity (Kagan, 1966) has been useful to describe differences in children's approaches

to problem solving. Whenever a number of response alternatives are simultaneously available and uncertainty as to the correct response is high, some children (reflectives) delay responding until the alternatives have been considered, and they have a high probability of being correct. In contrast, other children (impulsives) respond quickly with less thorough evaluation of the various possibilities and consequently make many mistakes. Using the latency and error measures of Kagan's (1966) Matching Familiar Figures (MFF) Test, children can be identified as cognitively reflective or impulsive.

This research was supported by grants from the National Institute of Mental Health (1 F31 MH-05270-01) and the Virginia Mental Health Fund to the first and second authors, respectively. Special thanks are due to James F. Hooke, who served as the sponsor for the first author's fellowship; to Donald J. Kiesler for the uniqueness of his encouragement; and to Shirley Truman for patient and responsible assistance.

At different times during the conduct and completion of the present article, the first author was affiliated with Virginia Commonwealth University, Virginia Treatment Center for Children, and the Palo Alto Veterans Administration Hospital.

Requests for reprints should be sent to Philip C. Kendall, Department of Psychology, Elliott Hall, 75 East River Road, University of Minnesota, Minneapolis, Minnesota 55455.

Among the most common behavior problems resulting in children being referred to mental health services is impulsiveness—that behavior pattern which involves a lack of inhibitory control and a tendency to respond quickly without thorough deliberation. In light of the severity of this behavior, an increasing number of studies have been directed at modifying impulsiveness in children (see Finch & Kendall, in press; Messer, 1976). Among the various methods that have been used in attempting to modify impulsivity are forced delay (Heider, 1971; Kagan, Pearson, & Welch, 1966), reinforcement contingencies

(Briggs, 1968; Finney, 1970), modeling (Debus, 1970; Denney, 1972), and instructions in strategies for scanning (Egeland, 1974; Nelson, 1969). In most of these studies, except where scanning strategies were emphasized, one or the other aspects of impulsivity (latency/errors) was modified but not both.

Other investigators (Finch, Wilkinson, Nelson, & Montgomery, 1975; Meichenbaum & Goodman, 1971) have examined verbal self-instructions training as a program for modifying impulsivity. These researchers, using cognitive training, reported changes in the desired direction for both latencies and errors.

The effects of reinforcement and response cost on cognitive style were investigated in a study by Nelson, Finch, and Hooke (1975), with results indicating that for impulsive subjects, the response-cost procedure was more effective. The utility of a response-cost procedure with impulsive children was supported by Erickson, Wyne, and Routh (1973).

Kendall and Finch (1976) combined the cognitive training procedures and the behavioral strategy of response cost into a cognitive-behavioral treatment and reported a successful case study in which both cognitive impulsivity and impulsive behavioral "switching" were reduced. In addition, the results of the cognitive-behavioral treatment generalized to the classroom and across a variety of situational tests. Even though this case study was suggestive of the utility of the cognitive-behavioral treatment, the clinical utility of such a treatment procedure needed to be demonstrated further using a clinic population in a group comparison study. The purpose of the present study was to conduct just such an examination of the effectiveness of the cognitive-behavioral approach in modifying impulsivity.

Method

Study Setting

The present study was conducted at the Virginia Treatment Center for Children, Richmond, Virginia, a university-affiliated children's psychiatric hospital. The hospital houses four living units, each with sleeping and play areas for 10 children and a respective classroom with two teachers.

Staff

Eight master's level special education teachers and 16 trained child care technicians provided ratings of behavior.

Subjects

The subjects for the present study came from all new admissions to the Virginia Treatment Center for Children between July 1, 1975, and March 1, 1976. The identification of impulsive children was based on their initial assessment scores on the MFF (Kagan, 1966). These scores, latency to first response and number of first-response errors, were compared to norms developed by us based on the performance of 195 children. The cutoff scores were as follows: Impulsives required an error rate ≥ 7 and a mean latency < 8.5 sec. Of the 51 children who were initially assessed, 20 children were identified as impulsive and were assigned to either a treatment ($n = 10$) or a control ($n = 10$) group according to a restricted randomization procedure.¹ The mean age of the children in the treatment group was 10.2 years and 11.1 years for the controls. The 20 impulsive subjects included 16 males and 4 females (8 and 2 in each group, respectively) of whom 4 were black and 16 were white (2 and 8 per group, respectively).²

Dependent Measures

Three types of dependent measures were examined: patient performance, self-report, and rating scales. Two self-report measures and three rating scales, two completed by the teachers and one by the unit personnel were used.

Patient performance. Each subject's performance on the MFF was examined at pretest, posttest, and follow-up periods. The MFF is a 12-item match-to-sample task that requires the child to choose from an array of six variants the one picture that is identical to a standard picture. This test assesses the conceptual-tempo dimension of reflection-impulsivity.

¹ The restricted randomization procedure allowed for random assignment of subjects to either the treatment or control group, with the restriction that not more than three subjects in a row be assigned to any one group. This was done to prevent having all treatment or all control subjects involved at the same time.

² Of the 20 subjects, there were 6 adjustment reactions of childhood or adolescence, 5 overanxious reactions, 5 neurotic conditions (3 depressive, 2 phobic), 2 aggressive reactions, and 2 psychotic organic brain syndrome. Relative dosages of Thorazine and/or Mellaril were prescribed for 30% of the subjects. There were no meaningful differences in the diagnoses or in the drugs for the two groups. In addition, 1 subject in each group had received Ritalin.

Latency to first response and first-response errors were recorded.

Self-report scales. The two self-report scales were the Impulsivity Scale (IS; Sutton-Smith & Rosenberg, 1959) and the Impulse Control Categorization Instrument (ICCI; Matsushima, 1964).

The IS is a 25-item (Hirschfield, 1965, revision) true-false self-report scale that assesses impulsivity defined as a tendency to be restless, to indulge in horseplay, to lose control, and to enter activities with excessive vigor. Test-retest reliability was .85 (Sutton-Smith & Rosenberg, 1959), and validity data (teacher ranking, IS scores, and classroom observation were significantly related) were acceptable (Hirschfield, 1965).

The ICCI contains 24 sentence situations to which subjects state degrees of choice between spontaneously impulsive-aggressive behavior and behavior requiring impulse control on a 4-point continuum. The scale assesses self-control over immediate action when aroused. Odd-even reliability was .93, and acceptable validity data according to an interview schedule and behavioral task persistence are provided (Matsushima, 1964).

Each child completed the inventories at the initial assessment (pretreatment), posttreatment, and follow-up periods.

Rating scales. The two rating scales used were the Impulsive Classroom Behavior Scale (ICBS; Weinreich, 1975), which was completed by the teachers, and the Locus of Conflict (LOC) Scale (Armstrong, 1971), which was completed by the unit personnel as well as the teachers.

The ICBS is a 9-item 5-point teacher rating scale that was developed by choosing the most frequently used descriptions for impulsive childhood behaviors taken from related text materials (e.g., attention span, work consistency). This scale was specifically developed for research involving the modification of impulsivity (Weinreich, 1975).

"Locus of conflict" refers to the predominant mode of impulse modulation that is exhibited by an individual. In internalization of conflict, the conflict is between impulses and their inhibition. Behaviors are rigidly controlled, and the individual experiences subjective discomfort. In contrast, externalization of conflict represents the conflict between the child's actions and the reactions that they bring about in others. Impulsive emotionally disturbed children are more likely to be rated as exhibiting externalization of conflict (Montgomery & Finch, 1975).

Each of the rating scales was completed by the respective rater at pretreatment, posttreatment, and follow-up periods.

Training Materials

There were six sets of training materials used in the present study, one for each of six therapy sessions. The six sets were (a) conceptual thinking, (b) attention to detail, (c) recognition of identities, (d) sequential recognition, (e) visual closure, and (f) visual-motor reproduction.

Set 1. This was a series of 48 plates containing four pictures, three of which were conceptually similar. The task was to "find the one that doesn't belong with the others."

Set 2. Four pictorial stimuli (in only a few cases there were nine) were presented. Two of the stimuli were identical, and the subjects were instructed to "find the pictures that match" (42 plates).

Set 3. This series of 191 plates was directed toward the recognition of identities under conditions of conceptual similarity. Each plate of two pictures required the child to verbalize, "Are these the same or are they different?"

Set 4. A series of figures (beads or other geometric figures) were presented in a sequential pattern. In each of the 68 sequences, the child was to choose from an array of alternatives "which one would come next."

Set 5. In this series of exercises, the child was presented an incomplete line drawing superimposed on a square configuration of evenly spaced dots. The child was given a pencil and was asked to "complete the drawing so that it looks symmetrical—the same on both sides. Just draw one line between any two dots." There were 50 of such visual closure tasks.

Set 6. This was a set of 56 visual-motor reproduction tasks in which the subject was asked to reproduce one design on a configuration of dots on a similar blank configuration of dots.

Period 1: Initial Assessment

All children entering the Virginia Training Center for Children were screened within a 3-day period, 10 days after their arrival at the Center. Almost without exception, children were taken from their respective classrooms between the hours of 9:00–11:30 a.m. or 1:00–2:30 p.m. Each child was initially given the opportunity to choose not to attend, but no one made such a choice. At times, however, scheduled sessions were delayed due to special events, misconduct, or illness.

During the initial assessment each child was administered the MFF, the IS, and the ICCI in a random order. Subjects were merely informed that the examiner was collecting some information and needed all the children to do these tasks. Subjects were also informed that their scores did not go to their therapists or teachers but that they were included in a larger group of scores. Children received a small candy reward for their cooperation.

Concomitant with the child's assessment session, the respective classroom teacher completed the ICBS and the LOC scale. Similarly, a randomly selected full-time unit person rated the child using the LOC scale. In order for unit ratings to be representative of unit behavior, each unit staff had to consult the records of the child's behavior on the unit for the period of time in question. Neither teachers nor unit persons were informed as to the assignment of subjects to groups, thus guaranteeing blind ratings.

Intervention

One male therapist was the individual trainer for both the treatment and control groups. All children received six 20-minute sessions during which they were exposed to the training materials and after which they, in some fashion, received a reward for their participation. Except for the instructions relating to the cognitive-behavioral treatment, subjects in both groups were given identical task instructions and performance feedback. No criterion number of tasks were set, rather, all subjects worked for 20 minutes. Thus, disregarding the cognitive-behavioral treatment, similar conditions were achieved for both subject groups.

Treatment group. Although this group received conditions similar to controls, these subjects also received specific additional training in verbal self-instructions and a response-cost procedure contingent on their errors during training.

1. Verbal self-instructions. In relation to each set of training materials, the verbal self-instructions were practiced in the following sequence: first, the therapist performed the task out loud, verbalizing out loud about possible answers and relevant aspects of the stimuli; second, the subject performed the task items with the instruction to talk out loud as the trainer had (the trainer provided guidance for overt self-instructions); next, the trainer performed the task while talking to himself in a whisper; then, the subject was told to "try the next item, and this time just whisper to yourself"; last, the therapist performed an item while modeling covert self-instructions, which, in turn, was followed by the child performing similarly.

The following sample of a self-instructional procedure was that used with the visual association tasks (Task Set 1).

Let's see now, What am I supposed to do? I'm supposed to find the one that doesn't belong with these others. I see four pictures here, so I better look at each one carefully. Okay, the first one is a clock, so is the second one. This one is a grandfather clock, but this one is a cup and saucer. So, I've got three clocks and one cup and saucer. It's the cup and saucer that doesn't belong.

In a later item done by the trainer, a planned error was made and corrected as follows: "Here we have all four animals. They're all animals . . . wait . . . this one isn't a dog, it's a lion. There, now I can correct myself before I make an error. The lion is the one that doesn't belong." Verbal self-instructions were performed during each of the six therapy sessions.

2. Response cost. At the beginning of each training session, the subject was given 10 chips and was told that the chips were his/her to keep but that he/she could lose 1 for making a mistake. An example was given, and it was checked to see that each subject understood the contingency. A reward menu hung over the work area, and on it were several categories of candy and gum purchasable with

9, 7, or 5 of the chips. Thus, it was explained to each child that he/she could use the chips to purchase a reward at the end of the session, and the more chips, the bigger and better the choice of reward. After each error on the training material, 1 chip was taken from the subject, and the inaccuracy was stated as the reason for the loss of a chip.

Control group. Although this group received similar conditions as did the treatment group, they did not receive verbal self-instructions nor was response cost contingent on errors. They were given their choice of rewards noncontingently (for their cooperation) at the end of each session.

Period 2: Posttreatment Evaluation

The posttreatment evaluation was conducted 4 weeks after the initial assessment. Subjects were told that all the children were taking the tests more than once. This was done to reduce the potential of a child perceiving the second test as due to failure on the test the first time. As in the initial assessment, subjects were taken from their respective classrooms to perform these tasks, and ratings were obtained from teachers and unit persons.

Period 3: Follow-up

Follow-up data were collected 3 months after the initial assessment (2 months after the posttreatment evaluation). Again, each child was individually administered the tasks and was told that all children were taking the tests several times. Ratings were again obtained from teachers and unit persons.

Results

Reliabilities

Patient performance. The reliability of the MFF performance was calculated using scores for all subjects who were initially assessed except those in the treatment group and those who were discharged during the course of the study ($n = 30$).³ The test-retest correlations for latency and errors at the different administrations were .72 and .78, .79 and .69, and .82 and .62 (first and second administration, second and third, and first and third, respectively). All correlations were significant ($p < .001$). Also, MFF latency and error measures correlated $-.55$, $-.75$, and $-.73$, respectively, for the three periods.

³ This group of subjects included children who were assessed at all periods and for whom data were complete, but it did not include those subjects in the treatment group. The control group was included.

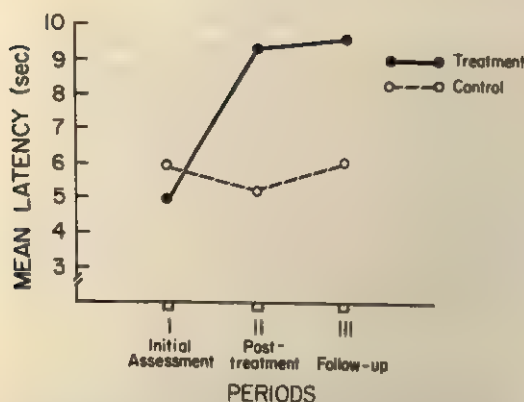


Figure 1. Mean latency in seconds for the treatment and control groups at the initial assessment, post-treatment, and follow-up periods.

Self-report scales. For the IS, test-retest reliabilities were .56 ($p < .005$), .70 ($p < .001$), and .33 ($.05 < p < .10$; first and second administrations, second and third, and first and third, respectively). In similar respective order, reliabilities for the ICCI were .56, .63, and .52 (all $ps < .005$).

Rating scales. The interrater reliability of each of the rating scales was determined using Pearson product-moment correlation coefficients. Reliabilities were assessed for teachers and unit personnel by assigning randomly selected children to each of a pair of raters. Each pair of raters was assigned five children, and there were six rater pairs.

The reliability of the ICBS was .85 ($p < .01$) for the sample of students rated by the teacher pairs. The reliabilities of the three measures from the LOC scale (internalization, externalization, and total maladjustment) were assessed for teacher and unit personnel pairs both combined and separately. The combined LOC scale reliabilities were .62, .93, .91 ($ps < .005$) for the three LOC scale measures, respectively. The teacher LOC scale reliabilities were .89, .96, and .95, and the unit personnel LOC scale reliabilities were .42, .82, and .84, in the same order ($n = 5$). Thus, the reliabilities of the rating scales were favorable.⁴

Group Comparisons

To evaluate changes in the dependent measure for subjects in the treatment and con-

trol groups across the three periods, separate 2 (between-subjects) \times 3 (within-subjects) analyses of variance were conducted.⁵

Patient performance. Separate 2 \times 3 analyses of variance were conducted for latency and error measures. The latency analysis resulted in a significant groups effect, a significant periods effect, and a significant Groups \times Periods interaction, $F(1, 18) = 7.02$, $p < .01$, $F(2, 36) = 9.17$, $p < .001$, $F(2, 36) = 11.87$, $p < .001$, respectively. These results are presented in Figure 1. An analysis of simple effects (via independent t tests)⁶ indicated that although the treatment and control groups' latencies did not differ at the initial assessment period, $t(18) = .96$, the groups differed significantly at both posttreatment and follow-up, $t(18) = 3.62$, $p < .001$, $t(18) = 2.58$, $p < .02$, respectively. In addition, a related t test indicated that for the treatment group, the change in latency from initial assessment to posttreatment was significant, $t(9) = 3.69$, $p < .01$. All other t tests were nonsignificant.

The analysis of variance of the error data resulted in a significant groups effect, a significant periods effect, and a significant Groups \times Periods interaction, $F(1, 18) = 13.67$, $p < .002$, $F(2, 36) = 17.30$, $p < .001$, $F(2, 36) = 6.45$, $p < .005$, respectively. These results are presented in Figure 2.

An analysis of simple effects (via independent t tests) indicated that although the treatment and control groups' error scores did not differ at the initial assessment periods, $t(18) = .37$, the groups differed significantly at posttreatment, $t(18) = 2.92$, $p < .01$, and the groups differed significantly at follow-up, $t(18) = 1.76$, $p < .05$. The decrease in the error rate for the treatment group from the

⁴ The intercorrelations of all dependent measures, in tabular form, are available from the first author.

⁵ A table of all means and standard deviations for all the dependent measures for the treatment and control groups is available from the first author.

⁶ In the cases in which specific a priori predictions were made, one-tailed t tests were used. As a point of information, however, all ts (except one) that were either significant or not significant for one-tailed tests would have been similarly significant or not significant if two-tailed tests had been used.

initial assessment to posttreatment was significant, $t(9) = 2.52$, $p < .05$, but the decrease from posttreatment to follow-up for the control group was not significant, $t(9) = 1.17$. All other t tests were nonsignificant.

Self-report scales. The results of the 2×3 analysis of variance for the IS resulted in nonsignificant main effects for groups and for periods ($F_s < 1$) and a nonsignificant Groups \times Periods interaction, $F(2, 36) = 2.37$, $p > .10$. Similarly, the ICCI analysis resulted in nonsignificant main effects and a nonsignificant interaction, $F(1, 18) < 1$, $F(2, 36) = 1.92$, $p > .10$, $F(2, 36) < 1$, respectively.

Rating scales. The results of the 2×3 analysis of the ICBS indicated nonsignificant main effects for groups and for periods ($F_s < 1$), but the Groups \times Periods interaction was significant, $F(2, 36) = 12.19$, $p < .001$. The nature of this interaction is presented in Figure 3. An analysis of simple effects via independent t tests indicated that the treatment and control groups' ICBS ratings differed at the initial assessment, $t(18) = 4.5$, $p < .001$, with the treatment group rated as more impulsive in the classroom. Although the groups only approached being significantly different at the posttreatment period, $t(18) = 1.74$, $p < .10$ (treatment group was rated as less impulsive), the groups differed significantly at follow-up, $t(18) = 2.85$, $p < .02$, with the treated subjects rated as less impulsive in the classroom. The change in ICBS ratings for the treatment group from initial assessment to posttreatment was a significant reduction, related $t(9) = 3.70$, $p < .01$. In addition, related t tests for the change in ICBS ratings of the control subjects indicated that the increase in impulsivity from initial assessment to posttreatment was significant, $t(9) = 2.61$, $p < .05$. The increase from posttreatment to follow-up was not significant, $t(9) = 1.43$.

Three 2×3 analyses of the LOC scale data were conducted for each of the raters (unit personnel and teachers) and for each of the three LOC scale measures (internalization, externalization, and total maladjustment). The results of the unit ratings of internalization were nonsignificant (all $F_s < 1$). Unit ratings of externalization of conflict resulted in a nonsignificant groups effect and

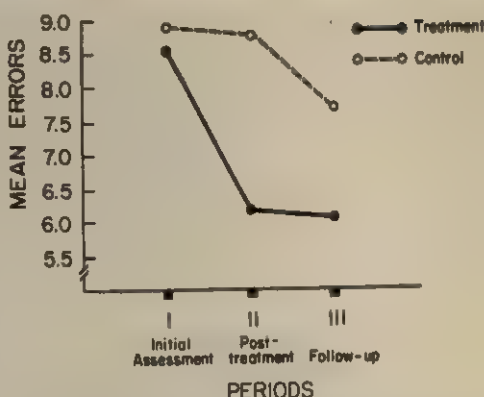


Figure 2. Mean errors for the treatment and control groups at the initial assessment, posttreatment, and follow-up periods.

a nonsignificant periods effect, $F(1, 18) < 1$, $F(2, 36) = 3.13$, $.05 < p < .10$, respectively. The Groups \times Periods interaction was significant, $F(2, 36) = 4.22$, $p < .02$. Analyses of the simple effects via related t tests for the treatment and control groups separately resulted in nonsignificant changes across periods for the treatment group, $t_s(9) < 1$, and a significant increase for the control group from the initial assessment to posttreatment, $t(9) = 2.88$, $p < .02$, but the posttreatment decrease to follow-up was not significant, $t(9) = 1.41$. Independent t tests indicated that although the groups differed significantly on the units' initial ratings of externalization, $t(18) = 2.62$, $p < .02$, they did not differ at

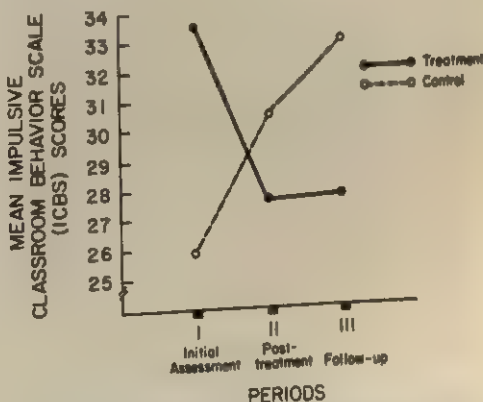


Figure 3. Mean Impulsive Classroom Behavior Scale (ICBS) scores for the treatment and control group at the initial assessment, posttreatment, and follow-up periods.

the posttreatment period, $t(18) < 1$, or at the follow-up period, $t(18) = 1.25$. Thus, the significance of the above analysis is attributable to a significantly lower externalization rating for controls at the initial assessment period; no other means differed significantly.

The total maladjustment ratings made by the unit personnel were analyzed using a 2×3 analysis of variance, which resulted in nonsignificant main effects for groups and for periods, $F(1, 18) = 1.10$, $F(2, 36) = 1.09$ (in that order), and a nonsignificant Groups \times Periods interaction ($F < 1$).

The teacher ratings of locus of conflict were similarly analyzed. The results of the analysis of the internalization data were nonsignificant main effects for groups and periods and a nonsignificant interaction, $F(1, 18) < 1$, $F(2, 36) = 2.43$, $F(2, 36) = 1.14$, all $ps > .10$ (in that order). The teacher externalization of conflict ratings produced a nonsignificant groups effect, $F(1, 18) = 1.16$, $p > .10$, a significant periods effect, $F(2, 36) = 3.63$, $p < .05$, and a nonsignificant Groups \times Periods interaction, $F(2, 36) = 1.18$, $p > .10$. Similarly, the teacher total maladjustment ratings produced a nonsignificant group effect and a nonsignificant interaction ($Fs < 1$), whereas the periods main effect was significant, $F(2, 36) = 4.49$, $p < .02$.

One additional question remained unanswered by the above analyses. The question was, Does the actual number of response-cost occurrences for subjects in the treatment group correlate with improvement? To examine this, the change from the initial assessment to the posttreatment period (improvement) for both MFF latencies and errors and for ICBS scores was correlated with the respective number of response-cost consequences occurring during treatment. The correlation between response-cost occurrences and improvement in MFF latency was not significant ($r = -.36$) nor was the correlation with MFF errors ($r = .36$). The correlation of response-cost occurrences and improvement on the ICBS was .58. This relationship ($p < .06$) suggests that the more response-cost occurrences, the greater the improvement in the classroom.

Discussion

The reliabilities of the latency and error measures of the MFF provided in the present study were in the moderate to high category, and they indicate, in contrast to Finch, Dearsdorff, and Montgomery (1974), that the measures are reliable over a 3-month period. Correspondingly, the reliability of the MFF performances suggests that the reflection-impulsivity dimension is relatively stable. The teacher ratings of impulsive classroom behavior (ICBS) were highly reliable. In addition, some validation information was provided in that the ICBS was found to be a sensitive measure of the effects of treatment. Along with the brevity and simplicity of format of the ICBS, the present study provided supportive reliability and validity data that should also be considered in selecting a measure of impulsive classroom behavior.⁷

The major results of the present study substantiate the effectiveness of a cognitive-behavioral program for the modification of impulsivity in emotionally disturbed children (Kendall, 1976). In addition to the alteration of both latency and error measures of cognitive tempo, the present study found a favorable generalization of treatment to the classroom. There, as evidenced by teacher ratings, the treatment subjects displayed significantly less impulsive classroom behavior. Although teacher ratings of locus of conflict did not reflect these differences, the present results support those of Kendall and Finch (1976), and, taken together, they are considered to support the utility of the cognitive-behavioral treatment for modifying impulsivity and attaining generalization to the classroom.

It should be noted that the desired generalization of the treatment effects to the classroom is in contrast with the increase in impulsivity of the control group. Although all subjects received the "constant" treatment milieu of the Virginia Treatment Center for Children, the increase in the control subjects' impulsiveness may be due to the dynamic theoretical model of the treatment center as a whole in which the expression of feelings

⁷ A copy of the ICBS can be obtained from the first author.

are emphasized. Even though this may or may not be the treatment model of choice for overly inhibited children, it would not appear to be the desired model for impulsive children. Moreover, these findings illustrate the efficacy of examining differential treatment effectiveness for distinct groups of children (Kendall & Finch, in press).

Some of the findings of the present study have general implications. The most outstanding of these is the fact that although neither of the self-report indicants of impulsivity changed due to treatment, two measures of MFF performance and ratings of classroom behavior reflected a desirable treatment effect. The implication here is that behavior change may occur without first altering self-perceptions. An equally compelling implication is that self-report scales may be of limited utility in treatment research with children (relatively insensitive to change).

In theorizing about the effectiveness of the cognitive-behavioral treatment, one must not ignore the training materials. Indeed, in the present study in which generalization to the classroom was attained, the tasks were of the psychoeducational variety. On the other hand, if the six treatment sessions consisted of cognitive training and response-cost dealing with interpersonal situations, attaining generalization to life situations may have been more likely. Thus, although it is not impossible to conclude that the cognitive-behavioral treatment did not generalize to the units, it is likely that the training tasks are relevant in regard to the type of generalization that will be obtained (see also Kendall, 1977).

Although the present study has demonstrated that the cognitive-behavioral treatment effectively modifies impulsivity, the treatment package contains many components—modeling, self-instructions, and response cost. Future research should examine the specific components of the treatment package in an attempt to isolate the one active component or, more likely, to uncover the relative efficacy of the components.

References

Armentrout, J. A. Parental child-rearing attitudes and preadolescents' problem behaviors. *Journal of*

- Consulting and Clinical Psychology*, 1971, 39, 276-285.
- Briggs, C. H. An experimental study of reflection impulsivity in children. Doctoral dissertation, University of Minnesota, 1969. Dissertation 48-17061, 1969, 24, 1491H. University Microfilms No. 68-1610.
- Debus, R. L. Effects of brief observation of model behavior on conceptual tempo of impulsive children. *Developmental Psychology*, 1971, 7, 77-82.
- Denney, D. R. Modeling effects upon conceptual style and cognitive tempo. *Child Development*, 1972, 43, 108-119.
- Egeland, B. Training impulsive children in the use of more efficient learning techniques. *Child Development*, 1974, 45, 165-171.
- Errickson, E. A., Wyne, M. D., & Routh, D. R. A response-cost procedure for reduction of impulsive behavior of academically handicapped children. *Journal of Abnormal Child Psychology*, 1971, 1, 350-357.
- Finch, A. J., Jr., Deardorff, P. A., & Montgomery, I. E. Reflection impulsivity: Reliability of the Matching Familiar Figures Test with emotionally disturbed children. *Psychological Reports*, 1976, 39, 1133-1134.
- Finch, A. J., Jr., & Kendall, P. C. Impulsivity: From research to treatment. In A. J. Finch, Jr., & P. C. Kendall (Eds.), *Treatment and research in child psychopathology*. Hohenwood, NY: Spectrum Publications, in press.
- Finch, A. J., Jr., Wilkinson, M. D., Nelson, W. M. III, & Montgomery, I. E. Modification of an impulsive cognitive tempo in emotionally disturbed boys. *Journal of Abnormal Child Psychology*, 1975, 1, 47-51.
- Finney, B. J. The modification of conceptual tempo in disadvantaged boys. Doctoral dissertation, Western Reserve University, 1968. Dissertation Abstracts International, 1969, 10, 1962A. University Microfilms No. 68-1611.
- Heider, E. R. Information processing and the modification of an "impulsive" conceptual tempo. *Child Development*, 1971, 42, 1774-1781.
- Hirschfeld, P. P. Response set in impulsive children. *Journal of Genetic Psychology*, 1967, 107, 11-13.
- Kagan, J. Reflection impulsivity: The generality and dynamics of conceptual tempo. *Journal of Abnormal Psychology*, 1968, 71, 1-14.
- Kagan, J., Pearson, I., & Wex, L. Modifiability of an impulsive tempo. *Journal of Educational Psychology*, 1968, 60, 450-453.
- Kendall, P. C. (Producer). *A cognitive-behavioral treatment for impulsive Minnagood*. University of Minnesota, 1970. Film.
- Kendall, P. C. On the effectiveness of verbal self-instructional procedures with children. *Cognitive Therapy and Research*, 1977, 1, 171-181.
- Kendall, P. C., & Finch, A. J., Jr. A cognitive-behavioral treatment for impulse control: A case study. *Journal of Consulting and Clinical Psychology*, 1970, 46, 852-857.
- Kendall, P. C., & Finch, A. J., Jr. Strategies for re-

- search in child psychopathology. In A. J. Finch, Jr., & P. C. Kendall (Eds.), *Treatment and research in child psychopathology*. Hollingswood, N.Y.: Spectrum Publications, in press.
- Matsushima, J. An instrument for classifying impulse control among boys. *Journal of Consulting Psychology*, 1964, 28, 87-90.
- Meichenbaum, D. H., & Goodman, J. Training impulsive children to talk to themselves: A means of developing self-control. *Journal of Abnormal Psychology*, 1971, 77, 115-126.
- Messer, S. B. Reflection-impulsivity: A review. *Psychological Bulletin*, 1976, 83, 1026-1052.
- Montgomery, L. E., & Finch, A. J., Jr. Reflection-impulsivity and locus of conflict in emotionally disturbed children. *Journal of Genetic Psychology*, 1975, 126, 89-91.
- Nelson, T. F. The effects of training in attention deployment on observing behavior in reflective and impulsive children. (Doctoral dissertation, University of Minnesota, 1969), *Dissertation Abstracts*, 1969, 27, 2659B. (University Microfilms No. 68-17, 703)
- Nelson, W. M., III, Finch, A. J., Jr., & Hooke, J. F. Effects of reinforcement and response-cost on cognitive style in emotionally disturbed boys. *Journal of Abnormal Psychology*, 1975, 84, 426-428.
- Sutton-Smith, B., & Rosenberg, B. G. A scale to identify impulsive behavior in children. *Journal of Genetic Psychology*, 1959, 95, 211-216.
- Weinreich, R. J. *Inducing reflective thinking in impulsive, emotionally disturbed children*. Unpublished master's thesis, Virginia Commonwealth University, 1975.

Received April 12, 1977 ■

Influence of Contextual Cues on the Efficacy of Desensitization and a Credible Placebo in Alleviating Public Speaking Anxiety

Jeffrey M. Slutsky
Roosevelt University

George J. Allen
University of Connecticut

This investigation was designed to determine the extent to which contextual cues mediated the effectiveness of systematic desensitization and a plausible placebo in alleviating public speaking anxiety. After participating in a public speaking situation that allowed the collection of self-report, physiological, and behavioral manifestations of anxiety, 67 subjects were randomly assigned to receive five sessions of either desensitization, "T scope" therapy, or no treatment. Each of these conditions was conducted in a context that either stressed the clinical relevance of the procedure or presented the procedure as a laboratory investigation of fear without therapeutic implications. Analysis of changes both between groups and within individuals indicated that desensitization reduced public speaking anxiety in both contexts, whereas the placebo was effective only in the therapeutic setting. The superiority of desensitization was most pronounced on the physiological variables. The results are interpreted as indicating support for a counterconditioning, rather than an expectancy, interpretation of desensitization.

Early documentations of the effectiveness of systematic desensitization (e.g., Lang, Lazovik, & Reynolds, 1965; Paul, 1966) were followed by numerous investigations that have sought to isolate the active therapeutic ingredients of this procedure. One critical area that has received much recent attention is the influence of cognitive factors on the outcomes engendered by desensitization. Three general research strategies have generally been used in investigating this complex theoretical issue.

The first involves comparing desensitization against various placebo manipulations. The outcomes of this strategy have been inconsistent. Paul (1966) and Davison (1968) demonstrated the superiority of desensitization with evidence that supported a counterconditioning interpretation. Other investigators, however, have failed to differentiate

the effects of desensitization and placebos in alleviating test anxiety (Allen, 1971) or small-animal phobias (McReynolds, Barnes, Brooks, & Rehaven, 1973; Tori & Worell, 1973). This inconsistency may be partly due to differences in the perceived credibility of the treatment and placebo procedures that have been used (Borkovec & Nau, 1972).

A variant of this first strategy involves providing instructional sets to augment the efficacy of desensitization within a therapeutic context. In general, the results produced within this framework suggest that such manipulations yield very modest effects (McGlynn & Mapp, 1970; Woy & Efran, 1972).

The third and potentially most useful line of investigation involves presenting desensitization as either a therapeutically relevant treatment or a nonclinical laboratory procedure by means of instructional and contextual cues. Utilizing this paradigm, several studies (Miller, 1972; Rosen, 1974) have indicated that desensitization is more effective when presented as a clinically useful technique. In

Requests for reprints should be sent to Jeffrey M. Slutsky, Department of Psychology, Roosevelt University, 430 South Michigan Avenue, Chicago, Illinois 60605.

addition, Tori and Worell (1973) found no differences between desensitization conducted in a laboratory context and a plausible placebo procedure.

Several systematic influences run through these three studies that limit their adequacy in providing a crucial test of the competing theoretical models of desensitization. First, both Miller and Rosen reported that a number of subjects in their laboratory conditions evidenced awareness of the therapeutic qualities of desensitization. More importantly, all three investigations examined small-animal phobias, which lessens their relevance for two reasons. Bernstein (1973) has clearly demonstrated the influence of contextual cues in both the selection of animal phobias and their response patterns on self-report and behavioral avoidance measures. Borkovec (in press) provides impressive evidence indicating that exposure to fearful animal stimuli does not evoke extreme physiological activation. These researchers advocate the use of various interpersonal anxieties in analogue investigations of desensitization.

The present study examines the efficacy of desensitization and a highly credible placebo procedure administered within separate clinical and laboratory contexts in alleviating public speaking anxiety. The first and third research strategies were combined to clarify the extent to which contextual cues influence the specific and nonspecific effects of these procedures and to provide a test of the two dominant models of desensitization. Specifically, demonstrating that desensitization is superior to a placebo in reducing self-report, behavioral, and especially physiological indices of anxiety across contexts would support a counterconditioning explanation, whereas interactions between treatments and settings would argue for a cognitive-expectancy interpretation.

Method

Subjects

Of slightly over 1,500 students in an introductory psychology course who were offered the opportunity to participate in a program designed to improve public speaking, 116 expressed interest and 75 met several selection criteria including acknowledging a

high level of fear of public speaking on Geer's (1965) Fear Survey Schedule, attending a pretreatment screening session, and providing a monetary deposit as a guarantee that treatment sessions would be attended. Complete data were collected from a total of 67 subjects, 35 of whom participated in the clinic phase of the experiment. The remaining 32 subjects were told that they could not be accommodated in the therapy program because it was filled. Approximately 5 weeks later, after the therapy phase had been completed, these subjects were asked to take part in an ostensibly independent laboratory study of physiological concomitants of anxiety.

The deposits of these latter participants were returned when they were refused treatment, and they were later paid for their participation as research subjects. The participants averaged 18.6 years of age and were predominantly freshmen women.

General Experimental Design

Table 1 provides an overview of the general experimental design. The experiment was conducted in two separate contexts in two independent phases with different subjects studied in each. During the first phase, subjects assigned to the clinical setting were screened and received five 50-minute sessions of either desensitization or a placebo therapy administered by one of two advanced graduate students. Immediately after completion of this phase, the graduate students carried out identical treatment procedures in a research context with those subjects assigned to the laboratory setting. Assessment sessions were conducted before and after each phase, and two no-treatment groups were included. These control subjects were informed that they could not be accommodated in the program for which they were eligible, but they were asked to participate in the posttreatment assessment sessions. A debriefing session for all subjects was held following completion of the laboratory phase. Those subjects who received no treatment or some other ineffective treatment were then offered therapy. Thus, the basic design involved six independent groups in a complete factorial crossing of three treatment conditions in two contexts with repeated measurement before and after intervention. The therapeutic procedures were crossed with two therapists.

Process and Outcome Measures

The screening and posttreatment evaluation sessions closely followed Paul's (1966) procedures. Before these assessments, a measure of trait anxiety (Spielberger, Gorsuch, & Lushene, 1970) was obtained. After the program had been explained, participants were required to give a 2.5-min speech on an assigned topic. Immediately before the speech, each subject completed the Anxiety Differential (Husek & Alexander, 1963) and a questionnaire

form, devised by the present authors, which provided indications of state anxiety and somatic/emotional turmoil, respectively. Two physiological measures, a 15-sec pulse rate and a palmar sweat print, were also collected at this time. During the speech, two trained observers noted behavioral manifestations of anxiety on Paul's (1966) Timed Behavioral Checklist. Following the speech, which was given before an audience of 7-10, each subject completed the Personal Report of Confidence as a Speaker, as modified by Paul (1966).

Following each of the treatment sessions, subjects rated their perceptions of the interaction along seven 7-point semantic differential scales (e.g., tense-relaxed, fast-slow). Ratings of relaxation during image visualization, anxiety evoked by images, and clarity of images presented during the second through fifth sessions were made along 6-point dimensions. A termination questionnaire was also administered to assess subject involvement in the program as well as perceptions of the therapist/experimenter on nine bipolar adjectives.

Experimental Groups

Subjects were randomly assigned to three conditions, each of which was replicated in two contexts. All treated participants were attached to a polygraph that provided prerecorded printouts, which were shown at the end of each session. Subjects in the clinical context were told that the printouts indicated that they were responding well to treatment, whereas in the laboratory context, this feedback was interpreted as indicating support for the research hypotheses. Thus, even though participants were led to believe that they were either "good" clients or good subjects, positive expectations of the *therapeutic efficacy* of the procedures were induced only in the clinical setting.

Desensitization/clinical context. Subjects assigned to this condition came to a departmental therapy clinic and were informed that they would be receiving a demonstrably effective form of therapy that would ameliorate their fear of public speaking. The treatment rationale described by Paul (1966) was presented, and his procedures were replicated with regard to training relaxation, constructing a hierarchy, and desensitization proper.

Desensitization/laboratory context. Treatment was conducted in a laboratory setting. Following Leitenberg, Agras, Barlow, and Oliveau (1969), subjects were told that the experimenter was interested in studying the visualization of anxiety-producing images and concomitant physiological reactions. Relaxation was presented as a means of facilitating visualization. The actual treatment procedures paralleled those used in the clinical setting.

Placebo/clinical context. This placebo manipulation replicated the procedures used by Tori and Worell (1973). Subjects learned that they would be presented with tachistoscopic scenes of individuals gesticulating while speaking. The rationale explained how viewing subliminal scenes while relaxed would

Table 1
General Experimental Design

Weeks of study	Context	
	Therapy	Research
4-5	Pretreatment assessment	Refused participation in therapy study
6-11	Treatment	
12-13	Posttreatment assessment	Recontacted for laboratory research
14-15		Pretreatment assessment
16-21		Treatment
22-23		Posttreatment assessment

Note. Subjects were selected during Weeks 1-3.

facilitate the acquisition of more adaptive public speaking skills. Irrelevant images were actually used.

Placebo/laboratory context. The same tachistoscopic procedures were used with the rationale stressing the experimenter's interest in the effects of subliminal exposure to stress-related images. It was intimated that such stimuli would have relatively minor impact.

No treatment/clinical context. This group participated in the assessment procedures before and after the therapy phase. Subjects were informed during the pretreatment assessment that they could not be accommodated in the current program and were promised therapy at a later date.

No treatment/laboratory context. After the screening session, these subjects were scheduled to return for a second evaluation at the end of the research program.

Results

Comparability of the Experimental Conditions

To ascertain whether any systematic pretreatment differences between the groups existed, seven Treatment \times Context \times Therapist analyses of variance were computed. Since only 4 of 49 F tests were significant and since they did not fall into a consistent pattern, the essential equivalence of the groups was assumed.

Reliability Considerations

A pair of raters independently matched each palmar sweat print against 15 specimens to provide quantifiable data. Interobserver reliabilities were .94 and .90 for judgments

Table 2

Means of Self-report, Physiological, and Behavioral Anxiety Indices Collected Before and After Intervention

Measure	Clinic						Laboratory					
	Desensi- tization		Placebo		Control		Desensi- tization		Placebo		Control	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Anxiety differential	75.5	62.6	72.5	60.4	66.3	64.9	69.6	62.9	72.6	70.2	67.1	65.4
Questionnaire form	29.8	21.2	32.5	21.0	25.9	23.8	24.7	18.9	25.0	22.6	24.3	21.2
Public report of confidence as a speaker	22.5	15.8	22.7	17.7	18.2	15.4	19.6	15.9	21.8	17.9	19.1	18.0
Trait anxiety	42.7	37.9	42.0	42.3	42.9	40.6	41.7	38.5	42.9	41.7	42.8	42.0
Pulse rate	84.1	78.0	78.8	82.0	85.0	84.7	80.1	76.0	84.4	84.1	84.3	84.3
Palmar sweat	7.5	7.0	9.1	9.7	7.7	7.9	7.9	6.8	8.9	8.4	7.7	7.2
Behavioral checklist	43.4	36.7	43.4	35.8	40.9	38.4	43.6	36.1	43.4	42.0	42.0	41.0

made before and after treatment. The usable score for each print was obtained by averaging the two ratings. Each subject's pulse was converted into beats per minute by multiplying by four. Three pairs of observers were trained to rate anxiety on the Timed Behavioral Checklist. Interrater reliabilities of total scores were .79, .87, and .86 for the observer dyads. The scores of both raters were summed to provide an indication of the behavioral manifestations of public speaking anxiety.

Overall Effectiveness of the Manipulations

Means of the major dependent variables before and after the experimental manipulations are presented in Table 2. Treatment \times Context \times Time analyses of variance indicated significant effects for repeated testing on the Anxiety Differential, the questionnaire form, the Personal Report of Confidence as a Speaker, and the Timed Behavioral Checklist. Treatment \times Time interactions were significant for state anxiety, $F(2, 61) = 3.55$, $p < .05$, the questionnaire form, $F(2, 61) = 3.18$, $p < .05$, and showed tendencies on the Personal Report of Confidence as a Speaker, $F(2, 61) = 2.87$, $p < .07$, and pulse rate, $F(2, 61) = 2.67$, $p < .08$. These interactions were largely due to greater improvement in the

desensitized subjects. Reliable Context \times Time interactions on the Anxiety Differential and the questionnaire form indicated that the clinic context was associated with greater anxiety reduction.

The reliability of improvement within each condition was assessed by correlated t tests, which are presented in Table 3.

Desensitization was the only treatment that produced significant improvement on self-reported, behaviorally observable, and, most importantly, physiologically measured anxiety. The placebo led to improved self-reported and behavioral functioning when conducted in a clinical context, but it was no more effective than the passage of time when administered under the guise of research. The no-treatment control groups generally showed little change.

Analysis of Individual Changes

To investigate the effects of the manipulations on the performance of individual subjects, residual change scores were computed by subtracting actual posttreatment scores from the predicted score determined by regression analysis. This procedure is analogous to an analysis of covariance in that it equates existing between-group differences and removes changes due to statistical regression. Reliable improvement or deterioration for

each subject was defined as having a residual change score either greater or less than 1.96 times the standard error of measurement for that variable (the 95% confidence interval). Table 4 presents the percentages of subjects in each treatment group who manifested significant improvement or deterioration on each variable as a result of participating in the experiment.

These data generally support the conclusions drawn from prior analyses. Subjects receiving desensitization in both contexts or the placebo in the clinical setting showed the highest rate of improvement on the self-report measures, with higher rates of deterioration being noted for the remaining groups. The extreme variability generated by the clinical placebo procedure is noteworthy. Although the improvement figures parallel those produced by desensitization, deterioration percentages were of a higher magnitude. Desensitization also produced the highest percentages of improvement and the smallest amount of deterioration on the physiological and behavioral measures.

Analysis of items measuring subjects' perceptions of the treatment sessions were non-significant with the exception of one reliable difference relating to the level of activity between the therapists. Responses to the termi-

nation questionnaire revealed that subjects in the desensitization groups believed their visualizations to be significantly clearer, $F(1, 33) = 50.17, p < .001$, and more anxiety evoking, $F(1, 33) = 12.50, p < .01$, than participants in the placebo conditions. In addition, subjects in the clinical context viewed their therapists as more active and organized and reported greater expectations for therapeutic improvement, $F(1, 33) = 25.36, p < .001$, than subjects in the research phase. These therapist differences did not affect performance on the outcome measures, however.

Discussion

Perhaps the most important conclusion to be drawn from this study is the superiority of desensitization in reducing self-reported and behaviorally measured anxiety and in lowering physiological arousal. The effects of this treatment were robust in terms of their insensitivity to contextual cues. Subjects who were desensitized in a research setting manifested therapeutically relevant gains, in spite of the fact that they viewed the procedure as having little utility in this regard. Setting factors, on the other hand, exercised a powerful effect on the outcomes produced by the placebo manipulation, with this procedure be-

Table 3
Within-Treatment Correlated t Values Between Pretreatment and Posttreatment Means for Anxiety Indices

Treatment	Measure							
	Anxiety differential	Questionnaire form	Trait anxiety	Public report of confidence as a speaker	Pulse rate	Palmar sweat	Behavioral checklist	df
Desensitization clinic	3.38**	3.33**	3.47**	6.62***	2.10*	.87	4.48**	10
Placebo clinic	2.18	3.84**	-.17	2.49**	-.68	-.63	3.23*	9
Control clinic	.70	1.50	.71	2.95*	.12	-.21	2.49*	13
Desensitization laboratory	4.30**	3.03*	1.55	2.13	1.80	5.81**	3.04*	9
Placebo laboratory	1.23	1.09	.45	2.07	.10	.59	.75	9
Control laboratory	.56	3.07*	.16	1.15	.04	.69	.70	11

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 4

Percentages of Subjects in Each Condition Showing Reliable Improvement and Deterioration in Performance

Measure	Clinic			Laboratory		
	Desensitization	Placebo	Control	Desensitization	Placebo	Control
Anxiety	55	60	29	30	20	33
differential	18	40	57	10	70	42
Confidence as	55	40	36	30	30	16
a speaker	9	40	21	40	60	42
Questionnaire	64	70	29	60	30	33
form	18	20	71	20	60	42
Trait	63	30	46	50	30	36
anxiety	9	40	46	20	30	36
Pulse	55	50	50	60	40	33
rate	36	30	36	10	30	50
Palmar	55	30	22	80	40	41
sweat	36	70	50	10	40	33
Behavioral	45	70	21	70	30	17
checklist	10	20	43	20	60	58

Note. The top figure of each entry and the bottom figure of each entry refer to improvement and deterioration respectively.

ing as efficacious as desensitization in a therapeutic context but not reliably better than no treatment in a research-related atmosphere.

These findings seem useful in placing some of the previously reviewed discrepancies into a coherent perspective. It appears that many of the studies that reported differences between desensitization and placebo treatments (e.g., Davison, 1968; Paul, 1966) used pseudo-therapeutic manipulations of low credibility (Borkovec & Nau, 1972). Highly plausible placebos, such as the one developed by Tori and Worell (1973), seem to engender positive outcomes that match those produced by desensitization but only when administered in clinical contexts. Failure to find differences between these two types of treatment in the Tori and Worell study might well have been due to contextual differences whereby desensitization was presented only as a laboratory procedure while the therapeutic relevance of the placebo was stressed.

Thus it appears that embedding placebo manipulations in a therapeutic context is a necessary but not a sufficient condition for insuring their effectiveness.

The theoretical import of these findings in-

volves the support that they provide for Wolpe's (1958) counterconditioning model of desensitization. The study provides further evidence that this therapeutic procedure has a potent effect on alleviating anxiety related to a clinically relevant problem and is relatively insensitive to cognitive influences. Modification of this conclusion awaits the resolution of many problems in the expectancy literature (Wilkins, 1973).

References

- Allen, G. J. The effectiveness of study counseling and desensitization in alleviating test anxiety in college students. *Journal of Abnormal Psychology*, 1971, 77, 282-289.
- Bernstein, D. A. Behavioral fear assessment: Anxiety or artifact? In H. Adams & P. Unikel (Eds.), *Issues and trends in behavior therapy*. Springfield, Ill.: Charles C Thomas, 1973.
- Borkovec, T. D. Investigations of fear and sleep disturbance: Methodological, measurement, and theoretical issues in therapy outcome research. In G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research*. New York: Plenum, in press.
- Borkovec, T. D., & Nau, S. D. Credibility of analogue therapy rationales. *Journal of Behavior Therapy and Experimental Psychiatry*, 1972, 3, 257-260.

- Davison, G. C. Systematic desensitization as a counterconditioning process. *Journal of Abnormal Psychology*, 1968, 73, 91-99.
- Geer, J. H. The development of a scale to measure fear. *Behaviour Research and Therapy*, 1965, 3, 45-53.
- Husek, T. R., & Alexander, S. The effectiveness of the Anxiety Differential in examination stress situations. *Educational and Psychological Measurement*, 1963, 23, 309-318.
- Lang, P. J., Lazovik, A. D., & Reynolds, D. J. Desensitization, suggestibility and pseudo-therapy. *Journal of Abnormal Psychology*, 1965, 70, 395-402.
- Leitenberg, H., Agras, W. S., Barlow, D. H., & Oliveau, D. C. Contributions of selective positive reinforcement and therapeutic instructions to desensitization. *Journal of Abnormal Psychology*, 1969, 74, 113-118.
- McGlynn, F. D., & Mapp, R. H. Systematic desensitization of snake-avoidance following three types of suggestion. *Behaviour Research and Therapy*, 1970, 8, 197-201.
- McReynolds, W. T., Barnes, R. R., Brooks, S., & Rehagen, N. J. The role of attention-placebo influences in the efficacy of systematic desensitization. *Journal of Consulting and Clinical Psychology*, 1973, 41, 86-92.
- Miller, S. B. The contribution of therapeutic instructions to systematic desensitization. *Behaviour Research and Therapy*, 1972, 10, 159-169.
- Paul, G. L. *Insight versus desensitization in psychotherapy*. Stanford, Calif.: Stanford University Press, 1966.
- Rosen, G. M. Therapy set: Its effects on subjects' involvement in systematic desensitization and treatment outcome. *Journal of Abnormal Psychology*, 1974, 83, 291-300.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *STAI manual for the State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Tori, C., & Worell, L. Reduction of human avoidant behavior: A comparison of counterconditioning, expectancy, and cognitive information approaches. *Journal of Consulting and Clinical Psychology*, 1973, 41, 269-278.
- Wilkins, W. Expectancy of therapeutic gain: An empirical and conceptual critique. *Journal of Consulting and Clinical Psychology*, 1973, 40, 69-77.
- Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford, Calif.: Stanford University Press, 1958.
- Woy, J. R., & Efran, J. S. Systematic desensitization and expectancy in the treatment of speaking anxiety. *Behaviour Research and Therapy*, 1972, 10, 43-49.

Received April 14, 1977 ■

A New Measure of Psychological Androgyny Based on the Personality Research Form

Juris I. Berzins, Martha A. Welling, and Robert E. Wetter
University of Kentucky

In line with recent reconceptualizations of psychological masculinity and femininity as independent dimensions, the present article describes the development and validation of the PRF ANDRO scale whose Masculinity and Femininity subscales were drawn, on the basis of theoretical definitions, from the item pool of Jackson's Personality Research Form. Using data obtained from over 2,000 college students, the subscales of the PRF ANDRO scale are shown to be independent, reliable, minimally related to socially desirable responding, and substantially related to the corresponding subscales of Bem's Sex Role Inventory (correlations between .50 and .65) and to major personality dimensions. Further evidence of construct validity is adduced from the score patterns in 18 different samples that include clinical and other noncollege groups. Since the PRF ANDRO scale can be scored from the answer sheets of the Personality Research Form, other investigators may reanalyze prior studies with particular regard to the proposition that high levels of Masculinity and Femininity, jointly denoting psychological androgyny, predict greater interpersonal competence and transsituational adaptability than do traditionally sex-typed role orientations.

In recent years, the conceptualization and measurement of sex roles have undergone radical changes (Bem, 1974, 1976; Block, 1973; Constantinople, 1973; Kaplan & Bean, 1976; Pleck, 1975; Spence, Helmreich, & Stapp, 1975). The once self-evident assumption that high levels of sex typing promote psychological adjustment has come under increasingly sharp theoretical and empirical attack (e.g., Bem, 1976; Bem & Bem, 1970); behaviors thought to be "natural" correlates of biological sex differences, to be exemplified by "appropriately" masculine men and femi-

nine women, have lost their stereotypic potency. Rather, members of the human liberation movement argue that depending on the situation, the behavioral repertoires of healthy adults of both sexes can encompass behaviors previously assigned to the sexes differentially. The concept of psychological androgyny—the integration of culturally masculine and feminine attributes into one's self-definition and behavior—has become ascendant in recent perspectives on mental health (e.g., Bem, 1976; Kaplan, 1976).

We wish to express our appreciation to Douglas N. Jackson and Research Psychologists Press for permitting us to reproduce items from the Personality Research Form in our research. We are also grateful for the normative data contributed by Joyce Cheney, Claudia Dispennette, Charles Kondo, Kerry McIntosh, David Mee Lee, Vincent Nerviano, Wesley Ross, Phillip Russell, Connie Weet, and others.

Requests for reprints, unpublished materials, and a manual for the use of the PRF ANDRO scale should be sent to Juris I. Berzins, Department of Psychology, University of Kentucky, Lexington, Kentucky 40506.

Traditional measures of psychological masculinity-femininity, however, have been predicated on the assumptions that masculinity and femininity define the endpoints of a single bipolar dimension (the sexes as "opposites") and that measures of this dimension are best constructed by aggregating various items that show large endorsement differences between the sexes or between persons of differing sexual preferences (Constantinople, 1973). Adherence to these assumptions has generated measures that create virtually nonoverlapping distributions of scores for men and women, with the implication that the majority of per-

sons who score within the range typical for their gender are "appropriately sex typed," whereas the few persons who score outside that range presumably exemplify sex role confusion, deviance, ambivalence, or maladjustment. But it is clear that defining masculinity as "not femininity" and femininity as "not masculinity" precludes characterizing persons as *both* feminine and masculine (androgynous) or, logically, as *neither*. To permit the latter sex role categories to emerge, at least two independent dimensions, one for masculinity and one for femininity, are needed. If psychological masculinity and femininity are construed as independent dimensions of human functioning (albeit dimensions whose components in many cultures have been regarded as differentially desirable in men and women), then one could characterize at least some individuals as *both* masculine and feminine, instrumental and expressive (Parsons & Bales, 1955), assertive and yielding (Bem, 1974), or agentic and communal (Bakan, 1966; Block, 1973; Carlson, 1971).

Three self-report measures that embody this dualistic conception of sex roles have already been published. The Bem Sex-Role Inventory (BSRI; Bem, 1974) contains separate Masculinity and Femininity subscales, each comprised of 20 adjectives (7-point response format) judged differentially desirable for men and women in our society. The BSRI has been used to select participants in experiments that have demonstrated the greater behavioral flexibility of androgynous persons, relative to sex-typed ones, across several situations (see Bem, 1976). Spence et al. (1975) developed separate male-valued and female-valued subscales of a Personal Attributes Questionnaire, with each subscale comprised of 18 or more bipolar adjectives (5-point format) representing attributes judged stereotypically masculine or feminine but considered desirable for both sexes. Heilbrun (1976) extracted Masculinity and Femininity subscales from an earlier bipolar composite index based on the Adjective Check List (Cosentino & Heilbrun, 1964); items had been selected originally by contrasting the endorsement proportions of men identified with masculine fathers and women identified with feminine

mothers. All three instruments use a quadripartite classification of persons—high masculine/high feminine (androgynous), high masculine/low feminine (masculine typed), low masculine/high feminine (feminine typed), and low masculine/low feminine (undifferentiated or indeterminate).

The present article describes the development and psychometric credentials of a fourth instrument. It was inspired by and drew its theoretical rationale from Bem's BSRI but was developed from the items of a widely used multiscale personality inventory, the Personality Research Form (PRF; Jackson, 1967). Although various aspects of the validity of the Masculinity and Femininity constructs embodied in the subscales of this new instrument, the PRF ANDRO scale, have already been investigated in several contexts (Kelly & Worell, 1976; Berzins, Note 1; Welling, Note 2; Wetter, Note 3; Woods, Note 4), the Masculinity and Femininity subscales of the BSRI are regarded as the principal criteria for initial assessment of convergent validity in this article.

Rationale for Item Selection

The rationale outlined by Bem (1974) in constructing the BSRI included provisions for (a) separate Masculinity and Femininity scales, (b) items selected on the basis of *sex-typed* desirability (e.g., in American society, a masculine characteristic should be judged more desirable for a man than for a woman), and (c) items with generally positive content. We chose to apply this rationale to the PRF item pool, since we had conducted extensive research with this instrument and, were scale development successful, accumulated data could be reanalyzed.

The PRF is a multitrait inventory based on Murray's need theory. Form AA of the PRF (Jackson, 1967) contains 20 content scales (Abasement, Achievement, Affiliation, Aggression, Autonomy, Change, Cognitive Structure, Defence, Dominance, Endurance, Exhibition, Harmavoidance, Impulsivity, Nurturance, Order, Play, Sentience, Social Recognition, Succorance, Understanding) and two validity scales (Infrequency and Desirabil-

ity). Each scale contains 20 items, 10 keyed true and 10 false. Reviews of the PRF attest to its unusually good reliability and validity (cf. Wiggins, 1973).

To select potential Masculinity and Femininity items from the PRF content scale item pool (400 items), the original PRF scale placement of the items was disregarded, but their content was evaluated for (a) consistency with Bem's rationale and (b) consistency with rationally derived abstract definitions of the main content themes of Bem's Masculinity and Femininity scales. In the latter regard, the BSRI Masculinity scale was held to depict a dominant-instrumental dimension comprised of themes of social-intellectual ascendancy, autonomy, and orientation toward risk; the BSRI Femininity scale was judged to revolve around a nurturant-expressive dimension, containing themes of nurturance, affiliative-expressive concerns, and self-subordination. PRF items consistent with the masculine themes consequently were selected if they appeared more desirable in men than in women, whereas items consistent with the feminine themes were selected if more desirable in women than in men. To facilitate some degree of control over acquiescent responding in the eventual scales, both true- and false-keyed items were considered. Sixty-four (32 masculine, 32 feminine) PRF items were provisionally selected in this manner by the first two authors. These items turned out to have been drawn from 16 of the 20 PRF content scales (no items from Change, Cognitive Structure, Order, or Sentience), with 9 Masculinity items drawn from the Dominance scale, 11 Femininity items from the Nurturance scale, and no more than 6 items from any other scale. In subsequent item analyses, the provisional scales were shortened to 29 Masculinity items (19 keyed true, 10 false) and 27 Femininity items (17 keyed true, 10 false).

Sex-Typed Desirability Ratings

To evaluate the consensuality of the authors' judgments regarding the sex-typed desirability of the final 56 items, a separate study was conducted (Wicher, Note 5). When

scoring direction is adjusted appropriately (false-keyed items reflected), Masculinity scale items should be judged more desirable "for a man" than "for a woman," whereas Femininity scale items should be judged oppositely. Using Bem's (1974) rating format (7-point scales ranging from "not at all desirable" to "extremely desirable"), 177 students in introductory psychology (61 men, 116 women) rated all items; 30 men and 57 women judged the items for a man; 31 men and 59 women judged them for a woman; no judge rated both target sexes. Each PRF item was presented to judges in its full form and was followed by the question, "In American society, how desirable is it for a MAN (alternately, WOMAN) to mark this item TRUE?" Judges were cautioned to make normative (descriptive) rather than prescriptive judgments.

A 2 (judge sex) \times 2 (target sex) analysis of variance disclosed that the mean desirability of Masculinity items was 5.35 when the target was a man but was 3.29 when the target was a woman, $F(1, 173) = 497.57, p < .0001$. The mean desirability of Femininity items, in turn, was 5.28 when the target was a woman but 3.58 when the target was a man, $F(1, 173) = 392.01, p < .0001$. In neither analysis were effects associated with judge sex or the Judge \times Target Sex interaction significant. Analysis of individual items showed that all 56 target sex differences were in the expected direction, with 53 differences significant beyond the .05 level and 50 beyond the .001 level.

Participants and Measures used in Psychometric Analyses

The main portion of this article is based on two samples. Sample 1 ($n = 1160$) was comprised largely of students in introductory psychology at the University of Kentucky in 1974-1975. The mean age was about 20. Sample 1 members completed (a) the 85-item Interpersonal Disposition Inventory (IDI) comprised of the 56 items of the PRF ANDRO scale, the 20-item PRF Desirability scale, 5 items from the PRF Infrequency scale (if more than one of these was answered in the keyed direction, the respondent was excluded

Table 1
Normative Data and Sex Differences on the PRF ANDRO Scale in Two College Samples

Measure	Males		Females		F
	M	SD	M	SD	
Sample 1					
Masculinity	16.70	4.33	12.86	4.76	192.29*
Femininity	14.29	3.59	17.90	3.50	287.01*
Sample 2					
Masculinity	16.18	4.78	11.86	4.84	195.19*
Femininity	14.31	3.57	18.37	3.60	312.29*

Note. Sample 1 is comprised of 457 men and 703 women; Sample 2, of 434 men and 552 women.

* $p < .0001$.

from analyses) and 4 fillers and (b) the 60-item BSRI. During the second semester, participants also completed (a) a personal attitudes survey, which included, in alternating order, the 33 items of the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1964) and the 33 items of a self-esteem scale designed especially for this study as a measure of generalized self-evaluation (Welling, 1976; Wetter, Note 3) and (b) Rotter's Internal-External Locus of Control Scale (Rotter, 1966).

Sample 2 ($n = 986$) consisted of introductory psychology students tested in 1970. Mean age was 18.7. Their responses to the full PRF (Form AA) were scored for the Masculinity and Femininity subscales of the PRF ANDRO scale. It should be noted that the order of appearance and the context of PRF ANDRO scale items differ in the regular PRF and the IDI.

Normative Data

Table 1 presents normative data and tests for sex differences for the PRF Masculinity and Femininity scales when scored from the IDI (Sample 1) and the full PRF (Sample 2). Although sex differences in item endorsement had played no part in item selection, in both samples men exceeded women on the Masculinity scale and women exceeded men on the Femininity scale (both p s $< .0001$); the differences approximate or exceed one standard deviation in each instance.¹ Analysis of individual items revealed that 83.9% (47/56) of these showed significant ($p < .05$)

differences in the expected direction; no items were significant in the opposite direction. As one would expect if most individuals were sex typed in our culture, men also scored higher on the Masculinity scale than on the Femininity scale, and women showed an even greater difference in the opposite direction.

Although Table 1 also reveals some differences between the two samples (e.g., Sample 1 women scored higher than Sample 2 women on Masculinity), such differences may arise with equal plausibility from differences in age or academic level (Sample 1 is older), the different contexts in which items were embedded (IDI vs. PRF), or cultural changes (1970 vs. 1974-1975).

Scale Independence and Reliability

With regard to the theoretical requirement that Masculinity and Femininity scores be orthogonal (rather than strongly and negatively correlated), the correlations between the two scales in Samples 1 and 2, respectively, were $-.05$ and $-.11$ for men and $-.16$ and $-.24$ for women.

Estimates of the retest stability of scale scores, using an interval of approximately 3

¹ Users of the BSRI should note that on the BSRI Masculinity scale, the 452 men obtained a $M = 5.07$, $SD = .65$; the 703 women, $M = 4.54$, $SD = .69$; $F(1, 1153) = 172.66$, $p < .0001$. On the BSRI Femininity scale, men had a $M = 4.56$, $SD = .53$; women, $M = 5.14$, $SD = .53$; $F(1, 1153) = 322.09$, $p < .0001$. Alpha coefficients were .86 for Masculinity and .80 for Femininity.

weeks, were obtained from a separate sample of 55 men and 82 women enrolled in the first author's Personality classes; these students were unaware of our interest in sex roles at the time of testing. Rather than completing the same instrument twice, about half of the students completed the entire PRF in class and the IDI at home; the other half reversed the sequence. Given the differing item contexts and conditions of administration, the obtained retest coefficients ($r = .81$ for both the Masculinity and Femininity scales) should be regarded as conservative.

Although the explicit multidimensional definitions of Masculinity and Femininity that were used tend to lower interitem correlations and the resulting scale homogeneities, item-total correlations in eight independent analyses (2 Scales \times 2 Samples \times 2 Sexes) ranged between .12 and .52 (all $ps < .01$), with median values ranging from .28 to .35. The alpha coefficients for Masculinity were .76 in Sample 1 and .79 in Sample 2. For Femininity, the coefficients were .67 in Sample 1 and .70 in Sample 2.

Sex Role Classification

Following the procedure used by Bem (1977) and Spence et al. (1975), we performed median splits on the distributions of scores on the Masculinity and Femininity scales separately, after combining the sexes. In this manner, Masculinity scores of 15 or greater and 14 or less were designated as high and low, respectively; the comparable Femininity scale values were 17 or greater and 16 or less. The categories created thereby were termed *androgynous* (high masculinity/high femininity), *masculine typed* (high masculinity/low femininity), *feminine typed* (low masculinity/high femininity), and *indeterminate* (low on both). When Samples 1 and 2 were pooled, the resulting classification percentages for men and women, respectively, were androgynous, 18.7% and 20.3%; masculine typed, 48.7% and 13.6%; feminine typed, 10.4% and 48.1%; indeterminate, 22.1% and 17.9%. Roughly speaking, then, 1 of every 2 persons is "appropriately" sex typed, 1 in 5 is androgynous, 1 in 5 is indeterminate, and only 1 in 10 is "cross-typed."

Content Dimensions in the BSRI and the PRF ANDRO Scale

Since a rational analysis of the content themes of the BSRI gave us the inspiration to develop the PRF ANDRO scale, it is appropriate to examine empirically the extent to which such themes emerged in the self-descriptions of Sample 1 participants. Using the data of the 1,155 persons who had completed the BSRI, the 40 BSRI items and gender (male = 0, female = 1) were intercorrelated and factored by the method of principal components (unities in the diagonal), followed by varimax rotation. Application of Cattell's scree test to the nine eigenvalues greater than 1.0 suggested that eight factors (56% of total variance) should be rotated.

In terms of saliently ($r > .50$) loading items (shown in parentheses), the four factors defined exclusively by BSRI Masculinity items refer to Social Ascendancy (acts as a leader, dominant, has leadership abilities, forceful, aggressive, assertive, strong personality), Autonomy (independent, self-sufficient, self-reliant, individualistic), Intellectual Ascendancy (defends own beliefs, willing to take a stand), and Physical Boldness (athletic, competitive). The three factors defined by BSRI Femininity items describe Nurturant Affiliation (tender, compassionate, warm, gentle, sympathetic, affectionate, understanding, sensitive to the needs of others, eager to soothe hurt feelings), Self-subordination (childlike, gullible, flatterable), and Introversion (soft-spoken, shy). The eighth factor refers only to respondents' gender (.88) and the items *feminine* (.85) and *masculine* (−.87), indicating that the latter items are synonymous with biological gender. Overall, these results offer surprisingly good support for our initial conceptual analyses.

A comparable analysis conducted with the 56 PRF ANDRO scale items (plus gender) yielded 18 eigenvalues greater than 1.0.² Ap-

² The larger number of eigenvalues for the PRF ANDRO scale, as compared to the BSRI, is a function of the larger number of items, the lesser item reliabilities associated with true-false versus multi-point scales, and possibly the lack of balancing for acquiescence in the BSRI.

plication of the scree test suggested the advisability of rotating only eight factors (34% of variance), of which seven were interpreted.

The four factors defined exclusively by PRF Masculinity items referred to Social-Intellectual Ascendancy (e.g., "I seek out positions of authority"), Autonomy (e.g., "If I have a problem, I like to work it out alone"), Orientation Toward Risk (e.g., "I avoid some sports and hobbies because of their dangerous nature," keyed false), and Individualism (e.g., "I don't care if my clothes are unstylish, as long as I like them"). The two factors defined exclusively by PRF Femininity items addressed Nurturance (e.g., "When I see a baby, I often ask to hold him"), and a blend of Affiliative Concerns and Self-subordination (e.g., "I like to be with people who assume a protective attitude toward me"). The former (Nurturance) dimension showed the highest loading for gender (.70) among the factors. The seventh factor included three Femininity items (e.g., "I am usually the first to offer a helping hand when it is needed") and two Masculinity items (e.g., "When I see a new invention, I attempt to find out how it works"), collectively denoting Helpful Initiative.

Although both the BSRI and the PRF scales are clearly multidimensional, the similarities between the main themes of the respective Masculinity and Femininity subscales suggest that convergent validity coefficients should be appreciable.

Convergent and Discriminant Validity

To determine the convergences between the PRF and BSRI measures of Masculinity and Femininity as well as their respective independence from socially desirable responding (assessed by the PRF, BSRI, and Marlowe-Crowne Desirability scales), and to examine their relation to measures of self-esteem and beliefs in internal-external locus of reinforcement, the measures administered to Sample 1 were intercorrelated and are presented in Table 2.

Table 2.

The convergent validity coefficients for Masculinity were .60 for men and .65 for women; for Femininity, they were .52 and

Table 2
Correlations Between PRF and BSRI Sex Role Measures, Desirability Scales, Self-esteem, and Locus of Control

Measure	1	2	3	4	5	6	7	8	9
1. PRF Masculinity ^a									
2. PRF Femininity ^a	-.16***	-.05	.60***	-.05	.22***	.09*	-.00	.36***	-.15*
3. BSRI Masculinity ^b	.65***	-.10**	.05	.52***	.19***	.15**	.29**	.13	-.04
4. BSRI Femininity ^b	-.17***	.50***	-.03	.22***	.25***	.32***	.02	.39***	-.15*
5. PRF Desirability ^a	.14***	.15***	.27***	.17***	.19***	.29***	.15*	.07	-.04
6. BSRI Desirability ^b	.09**	.05	.21***	.22***	.47***	.50***	.30***	.55***	-.30***
7. M-C Desirability ^c	.14**	.18***	.11*	.23***	.40***	.37***	.33***	.38***	-.02
8. Self-esteem ^e	.38***	-.06	.44***	.11*	.60***	.54***	.36***	.18**	-.08
9. Focus of control ^d	-.08	.01	-.11*	-.02	-.33***	-.27***	-.27***	-.29***	-.16*

PRF = Personality Research Form; BSRI = Bem Sex-

Note. Coefficients for men and women are shown.

Role Inventory; M-C = Marlowe-Crowne.

* 457 men, 703 women.
 † 452 men, 703 women.
 ‡ 191 men, 339 women.

and 196 men, 369 women. High scores on this scale indicate a high level of self-esteem. *** $p < .001$. ** $p < .01$.

* $p < .05$.
** $p < .01$.

.50, respectively. (For the sexes combined, they were .68 and .61). These values indicate substantial similarities between the constructs underlying the total scores in these instruments. Analysis of individual items, in fact, showed that all PRF Masculinity scale items correlated significantly ($Mdn = .23$, all $ps < .001$) with BSRI Masculinity scores; similarly, all PRF Femininity items correlated significantly ($Mdn = .22$, all $ps < .04$ or better) with BSRI Femininity scores. Conversely, all BSRI Masculinity items and all but one BSRI Femininity item related significantly to their respective PRF counterpart scores (Masculinity $Mdn = .37$, Femininity $Mdn = .33$, all $ps < .02$ or better).

Excluding the BSRI items *masculine* and *feminine* (which refer to gender only), the high scorer on the PRF Masculinity scale is described by the 10 best BSRI items as acts as a leader, has leadership abilities, dominant, willing to take a stand, willing to take risks, independent, forceful, competitive, strong personality, and individualistic. The high scorer on the PRF Femininity scale, in turn, can be described as sympathetic, loves children, eager to soothe hurt feelings, sensitive to the needs of others, tender, compassionate, affectionate, gentle, warm, and understanding.

With regard to discriminant validity, the correlations of the sex role measures with the three Desirability scales (Measures 5, 6, and 7 in Table 2) are generally positive and low. For the two PRF scales, the 12 relevant coefficients range from .00 to .29; for the BSRI scales, they range from .02 to .32. For both instruments, consequently, socially desirable responding explains very little variance in scores.

Inspection of the correlations involving the Self-esteem scale, however, reveals that scores on both Masculinity scales are moderately related to favorable self-evaluations (for the PRF Masculinity scale, .36 among men and .38 among women; for the BSRI, .39 among men, .44 among women). In contrast, the relation of both Femininity scales to self-esteem is negligible, with coefficients ranging from $-.06$ to $.13$. Differences among sex role categories in self-esteem (Spence et al., 1975; Welling, 1976; Wetter, Note 3), consequently,

are more likely to be related to variations in Masculinity than in Femininity scores when the PRF ANDRO or the BSRI measures are used.

Finally, it is apparent that apart from low-order correlations suggesting a positive relation of Masculinity scores to beliefs in internal control, significant among men and for the BSRI among women, little variance is shared between these sex role measures and beliefs in internal-external control. As noted elsewhere (Berzins, Note 6), the best predictor of locus of control is the PRF Desirability scale, probably because it contains an appreciable self-esteem component (see Table 2) and has in fact been construed as a measure of self-esteem in a prior study (Berzins, Barnes, Cohen, & Ross, 1971).

Sex Role Categories and Personality Dimensions

Since the members of Sample 2 had completed the entire PRF, it was possible to examine the relation of their PRF ANDRO scale scores to the principal components of the 20 PRF scales. The latter were extracted in an earlier analysis (Berzins, Note 6), but they require brief description. Following standardization of PRF content scales within sex, the matrix was factored for the sexes separately and combined, with varimax rotation. In all three analyses, five components showed eigenvalues greater than 1.0 (explaining 64.4% of total variance in the combined analysis); between-sex congruency coefficients (Harman, 1967) ranged from .85 to .94; and invariance coefficients (Pinneau & Newhouse, 1964), from .75 to .93. In the combined analysis ($n = 986$), the components and their principal markers (loadings $> .60$) were Impulsivity (Impulsivity, .79; Cognitive Structure, $-.81$; Order, $-.76$); Dependency (Succorance, .86; Autonomy, $-.74$); Social Poise (Exhibition, .78; Affiliation, .76; Dominance, .65); Defensiveness (Defendence, .81; Aggression, .76; Abasement, $-.64$); and Intellectuality (Sentience, .80; Understanding, .74). Component scores ($M = 50$, $SD = 10$) derived from this solution by multiple regression should render less obtrusive the overlap

Table 3
Mean Factor Scores and Univariate F Ratios Associated with Four Sex Role Categories Across Five Principal PRF Components Among College Students

PRF component	Category				F
	Indeterminate	Feminine typed	Masculine typed	Androgynous	
Dependency					
Men	53.9 _a	61.2 _b	44.7 _c	51.8 _a	78.53***
Women	46.9 _a	55.9 _b	35.4 _c	45.6 _a	156.30***
Social Poise					
Men	41.4 _a	49.7 _b	51.5 _b	57.1 _c	51.16***
Women	43.2 _a	50.1 _b	48.3 _b	57.7 _c	44.57***
Intellectuality					
Men	44.5 _a	51.6 _{b,c}	50.3 _b	55.5 _c	20.05***
Women	44.5 _a	49.9 _b	51.6 _{b,c}	54.8 _c	20.31***
Defensiveness					
Men	49.0 _{a,b}	46.1 _b	51.6 _a	49.0 _{a,b}	5.25**
Women	50.7 _a	48.2 _a	56.2 _b	50.5 _a	12.38***
Impulsivity					
Men	51.2	47.0	50.5	49.3	2.45*
Women	53.3 _a	48.6 _b	53.7 _a	48.5 _b	9.46***
N					
Men	102	52	205	75	
Women	98	290	66	98	

Note. In each row, cells with the same subscript do not differ at the .05 level or better by Tukey's honestly significant difference test. All factor scores have been standardized to a mean of 50 and a standard deviation of 10. PRF = Personality Research Form.

* $p < .06$.

** $p < .002$.

*** $p < .0001$.

between PRF ANDRO scale items and the PRF scales from which they were drawn originally.

Differences between the four sex role groups, based on univariate analyses of variance computed for the sexes separately, are shown in Table 3. The PRF components are presented in order of decreasing intergroup differentiation.

Although the F ratios in 9 of the 10 rows of Table 3 reveal highly significant differences between groups, it is apparent that these differences are especially pronounced on the first two dimensions, Dependency and Social Poise. On the Dependency dimension, feminine-typed persons attained very high and masculine-typed persons very low scores; the difference between these two groups reached two standard deviations among women. Differences of this magnitude suggest that the median split procedure whereby these persons

were operationally defined as masculine typed (above the median in Masculinity, below in Femininity) and feminine typed (the reverse pattern) is functionally equivalent to splitting the distribution of Dependency component scores at its median and terming high scorers feminine and low scorers masculine! To assess this equivalence empirically, Androgyny difference scores (Femininity minus Masculinity; cf. Bem, 1974) were correlated with Dependency component scores. The resulting coefficients were .79 for men and .82 for women, suggesting that whenever Masculinity and Femininity scores differ greatly, as in strongly sex-typed persons, one might as well speak of individual differences along a dimension ranging from extreme dependency to extreme autonomy. Although the mean Dependency scores in Table 3 suggest some Gender \times Sex Role differences (e.g., indeterminate and androgynous men appear more dependent

than their female counterparts), the main thrust of these results "identifies" masculine versus feminine typing with variations in Dependency. Note that the two non-sex-typed groups, androgynous and indeterminate, differ from both sex-typed groups but not from each other.

On the Social Poise dimension, on the other hand, the pattern of means was very different. Here, androgynous persons scored much higher than did indeterminate persons, the mean difference between these groups being about $1\frac{1}{2}$ standard deviations. Androgynous individuals also scored significantly higher (and indeterminates lower) than the two sex-typed groups (which did not differ from each other). Since scores on the Social Poise dimension are orthogonal to scores on the Dependency dimension, differences in Social Poise obviously comprise an independent major dimension along which sex role groups vary. Since androgynous persons by definition have high scores on Masculinity and Femininity (whereas indeterminates have low scores), it is not surprising that Androgyny sum scores (Femininity plus Masculinity; cf. Strahan, 1975) should correlate .60 (for men) and .59 (for women) with scores on the Social Poise dimension.

Turning to intergroup differences on the Intellectuality dimension, the patterning of group differences was almost identical to that on Social Poise: Androgynous and indeterminate persons again defined the extremes, and androgynous persons scored significantly higher than at least one (and indeterminates lower than both) of the sex-typed groups. Differences on the Defensiveness dimension were less pronounced but resembled those associated with the Dependency dimension, in that among both sexes the masculine-typed persons were the most, and feminine-typed ones the least, defensive. Finally, differences on the Impulsivity dimension suggest that among women, feminine-typed and androgynous persons show greater impulse control than do masculine-typed and indeterminate persons. These differences did not reach significance among men.

Among college students, four of the five principal components have proven salient in

differentiating the groups. Although the resulting characterizations are technically four dimensional, two main patterns emerged: (a) Masculine-typed persons were the most independent and defensive, whereas feminine-typed persons were the most dependent and least defensive; (b) androgynous individuals were the most socially poised and intellectually oriented, whereas indeterminates were the most socially awkward and unintellectual. (These patterns have been fully cross-validated in samples of psychotherapists and alcoholics; see Berzins, Note 1.)

Broader Issues in Construct Validity

Thus far, this article has involved college students only. To facilitate an appraisal of the validity of the Masculinity and Femininity constructs across a variety of populations varying in age, education, occupation, and clinical status, Figure 1 arrays 18 different samples, representing over 6,000 persons, on both scales. Ten samples contain both men and women. Seven samples were scored from the PRF, 11 from the IDI.

The configuration of groups along the Masculinity (vertical) dimension strikingly corroborates the theoretical definition of Masculinity as an amalgam of social-intellectual ascendancy, autonomy, and risk taking. Policemen, men majoring in accounting, and male dental students showed the highest Masculinity mean scores; clinically depressed women, newlywed women, and women enrolled in a hospital weight-reduction program showed the lowest scores. Analyses conducted *within* groups (e.g., a sample of 1,109 college students) have also revealed interesting variations along the dominant-instrumental dimension. For example, college men who, 10 years from now, expected to make as much money as their classmates had mean Masculinity scores of 16.3; those who expected to make more scored 17.8; and those who expected to make much more scored 19.6, $F(2, 383) = 15.27$, $p < .0001$. The comparable values for college women were 12.7, 14.3, and 16.3, $F(2, 720) = 15.67$, $p < .0001$.

The theoretical alignment of Femininity with themes of nurturance, affiliative-expressive concerns, and self-subordination appears

to be corroborated by the fact that high school, college, and "general population" women showed high scores, whereas most male groups and, among women, gay women and women majoring in accounting scored considerably lower. It is curious, however, that both male and female psychotherapists showed relatively low Femininity (and Masculinity) scores in comparison to their respective gender peers. Perhaps because they are the most highly educated and testwise group sampled, and because their interpretation of some nurturance-related items may have been restricted to psychotherapeutic contexts, these therapists did not describe themselves more extremely on this dimension. Although no sample of unemployed housewives with children—who might be expected to show very high Femininity scores—is presented, analysis of college women's anticipated life-styles disclosed highly significant differences in Femininity scores as a function of the number of children these women anticipated having in their own future families. Women anticipating childless versus one-, two-, three-, and four- (or more) child families showed mean Femininity scores of 15.1, 17.2, 17.7, 18.6, and 18.9, respectively, $F(4, 718) = 17.68$, $p < .0001$. For college men, the corresponding values were 12.3, 15.0, 15.0, 14.6, and 14.7, $F(4, 381) = 3.18$, $p < .01$.

The overall distribution of mean scores in Figure 1 clearly shows that the vast majority of groups in our society are sex typed. If we were to treat the entries in Figure 1 as individuals rather than groups, median splits on both scales would classify 12 of 15 male groups as masculine typed and 11 of 13 female groups as feminine typed. The overall percentage (82%), of course, is much higher than that for individuals *within* any one of the groups. It does emphasize, however, the pervasive and highly significant sex differences on both scales in almost every sample tested to date. Indeed, only the self-designated homosexual volunteers did *not* show significant differences on these scales; both gay men and women emerged as quite high in Masculinity and moderate in Femininity (cf. Heilbrun & Thompson, 1977, for similar findings with the Heilbrun scale).

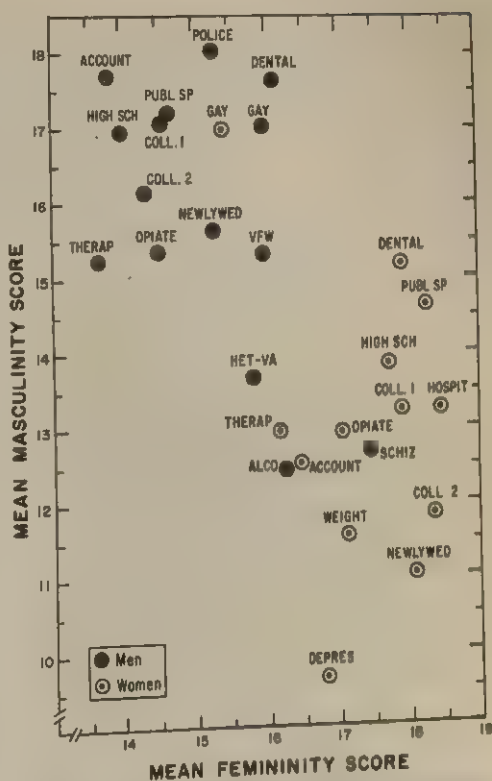


Figure 1. Distribution of Masculinity and Femininity scores across 18 samples (28 groups). (Account = college students in accounting classes, $n = 105$; Alco = hospitalized Veterans Administration alcoholics, $n = 760$; Coll. 1 and Coll. 2 = college students in introductory psychology classes in 1974-1976 and 1970, $n = 2,547$ and 986 , respectively; Dental = dental school students, $n = 54$; Depres = hospitalized clinically depressed patients, $n = 20$; Gay = homosexual volunteers, $n = 71$; Het-VA = heterogeneous nonschizophrenic Veterans Administration outpatients, $n = 33$; High Sch = high school students, $n = 700$; Hospit = hospitalized nonpsychiatric medical patients, $n = 20$; Newlywed = newlywed couples, $n = 198$; Opiate = civilly committed opiate addicts, $n = 216$; Police = municipal police, $n = 25$; Publ sp = college students in public speaking classes, $n = 166$; Schiz = schizophrenic Veterans Administration outpatients, $n = 20$; Therap = psychotherapists and trainees, $n = 276$; VFW = Veterans of Foreign Wars, $n = 57$; Weight = participants in a hospital weight-reducing program, $n = 54$.)

There are exceptions of clinical interest as well. Male schizophrenic outpatients and alcoholic inpatients, for example, showed markedly "feminine" typing. Within both sexes, psychopathology appears to be associated with normatively low Masculinity scores (e.g.,

clinically depressed female inpatients, heterogeneous nonschizophrenic male outpatients) and, among men but not women, with somewhat elevated Femininity scores. Whereas the problems that male schizophrenics have had with respect to gender identity are documented in the literature (cf. LaTorre, 1976), male alcoholics as a group have seldom been depicted as feminine typed by traditional measures of Masculinity-Femininity. From the present theoretical perspective, however, it is entirely reasonable to expect chronic alcoholics (M age = 43.9 years) to score very low on a dominant-instrumental dimension.

Whereas most of the groups in Figure 1 are relatively homogeneous in age (M ages range from 15.7 for high school students to 49.0 for the Veterans of Foreign Wars), cross-sectional comparisons among age groups in the more heterogeneous samples of men reveal low but significant decreases in Masculinity and/or increases in Femininity scores as age increases. Classification of subjects into the four sex role categories, consequently, should be based on normative considerations of age as well as those concerning education, socioeconomic status, and the degree of psychopathology.

General Discussion

The results suggest that the Masculinity and Femininity subscales of the PRF ANDRO scale are independent, reliable, minimally related to socially desirable responding, substantially related to their respective counterparts in the BSRI, convergent with major personality dimensions, and capable of meaningfully differentiating samples varying in age, socioeconomic status, and psychopathology. The appreciable convergences between the PRF ANDRO scale and the BSRI moreover suggest that experimental studies using the former to classify subjects should yield results similar to, if not identical with, studies using the latter (e.g., Bem, 1975; Bem & Lenney, 1976). Furthermore, the fact that the PRF ANDRO scale can be scored from the PRF answer sheets enables other investigators to reanalyze prior studies in terms of the sex role categories delineated in this study.

Substantively, the construal of Masculinity and Femininity as independent dimensions

rather than as opposites, and the resulting quadripartite classification of sex roles, has afforded a rather comprehensive view of the relation of sex roles to personality dimensions. Although masculine- and feminine-typed persons indeed emerged as opposites on some dimensions (e.g., Dependency, Defensiveness), they were relatively indistinguishable on others. On these other dimensions (e.g., Social Poise, Intellectuality), it was the androgynous and indeterminate persons who emerged as "polar opposites." Had the PRF ANDRO scale been constructed in the traditional manner, that is, by selecting PRF items that showed the largest sex differences in endorsement, it is obvious that neither the personality patterns of androgynous persons nor the sex-role-related differences in Social Poise and Intellectuality would have been detected.

To further appreciate the extent to which the present approach, like that of Bem's, differs from traditional ones, recall that one index of the "validity" of traditional measures has been the point-biserial correlation between gender and a measure of masculinity-femininity; such coefficients often exceed .70 (Constantinople, 1973). By this standard, both the PRF and BSRI scales did not fare very well: In Sample 1, gender correlated $-.38$ and $-.36$ with the PRF and BSRI Masculinity scales and $.45$ and $.47$ with the respective Femininity scales. But although sex differences may characterize both traditional and new measures of sex roles, their presence has little to do with overlapping content domains. For example, in a subset of Sample 2 participants ($n = 682$), the correlations between the (bipolar) Masculinity-Femininity scale of the Omnibus Personality Inventory (Heist & Yonge, 1968) and the PRF Masculinity scale were $.04$ (men), $.08$ (women), and $.30$ (combined); the corresponding coefficients with the PRF Femininity scale were $-.22$, $-.16$, and $-.42$. Apart from the higher coefficients resulting from sex differences in the combined sample, it is clear that our measures share trivial content variance with this traditionally constructed measure. In another sample of 206 male alcoholics, the correlations between the Minnesota Multiphasic Personality Inventory Mf scale and our

Masculinity and Femininity scales were $-.06$ and $.00$, respectively. Since our Masculinity and Femininity scales are independent rather than negatively correlated among both men and women, the pervasive sex differences observed in most groups (cf. Figure 1) are better understood as reflecting the consequences of traditional socialization practices (encouraging boys toward self-definitions thought more desirable in men than in women, vice versa for girls) than the psychometric "rediscovery" of biological gender accomplished by compounding items on which the sexes show differing endorsement frequencies.

The multithematic definitions of our Masculinity and Femininity constructs—contrasting social-intellectual ascendancy, autonomy, and risk taking with nurturance, affiliative-expressive concerns, and self-subordination—were corroborated empirically by the factor analyses of the BSRI and PRF ANDRO scale item pools. The reader should note that the "discovery" of these content themes in no way precludes using these instruments in their present form, nor does it argue for the utility of fractionating the present scales into subscales; all items in both instruments correlated significantly with their respective total scores and are intended to emphasize bandwidth over fidelity. Because of their breadth, however, the Masculinity and Femininity constructs can be amalgamated with related major conceptions in personality and interpersonal theory, for example, Bakan's (1966) distinction between agency and communion as male and female principles that require integration (see also Block, 1973; Carlson, 1971), Jung's (1956) animus/anima dialectic, Parsons and Bales's (1955) conception of instrumental and expressive roles, Leary's (1957) and Carson's (1969) organizing interpersonal behaviors around dominance-submission and love-hate coordinates (cf. Berzins, Welling, & Wetter, Note 7), and even Campbell's (1975) recent discussion of the tensions along a selfishness versus altruism dimension throughout history.

Overall, the definition of psychological androgyny that has emerged from the present study emphasizes a relative balance of masculine- and feminine-typed attributes in the context of high social competence (openness

to interpersonal and intellectual experiences). Block (1973) and others have marshaled evidence in support of the proposition that the attainment of higher levels of moral development is coextensive with the integration of traditionally masculine and feminine domains of self-definition and behavior. Conversely, individuals who are neither androgynous nor sex typed—indeterminate persons—are characterized by lower levels of social competence in this study and low self-esteem in others (e.g., Spence et al., 1975; Welling, 1976; Wetter, Note 3).

Finally, the quadripartite classification of sex roles, insofar as it converges conceptually and empirically with the dimensions of the interpersonal behavior circle (Carson, 1969; Leary, 1957; Berzins et al., Note 7) requires examination in the context of interpersonal communication, compatibility, and influence, especially in psychotherapy (Berzins, in press). Even though most therapists are men and most patients women, little research has been devoted to the appraisal of the conditions under which gender- and sex-role matching variables affect the processes and outcomes of psychotherapy. The ascendancy of psychological androgyny as a model of mental health suggests that such research has high priority.

Reference Notes

1. Berzins, J. I. New perspectives on sex roles and personality dimensions. In R. L. Bednar (Chair), *Sex roles: Masculine, feminine, androgynous, or none of the above?* Symposium presented at the meeting of the American Psychological Association, Chicago, August–September 1975.
2. Welling, M. A. A new androgyny measure derived from the Personality Research Form. In R. L. Bednar (Chair), *Sex roles: Masculine, feminine, androgynous, or none of the above?* Symposium presented at the meeting of the American Psychological Association, Chicago, August–September 1975.
3. Wetter, R. E. Levels of self-esteem associated with four sex role categories. In R. L. Bednar (Chair), *Sex roles: Masculine, feminine, androgynous, or none of the above?* Symposium presented at the meeting of the American Psychological Association, Chicago, August–September 1975.
4. Woods, M. M. The relation of sex role categories to autobiographical factors. In R. L. Bednar (Chair), *Sex roles: Masculine, feminine, androgynous, or none of the above?* Symposium presented at the meeting of the American Psycho-

logical Association, Chicago, August-September 1975.

5. Wicher, D. *Sex-typed desirability of PRF ANDRO scale items*. Unpublished manuscript, 1976. (Available from Donna Wicher, Department of Psychology, University of Kentucky, Lexington, Kentucky 40506.)
6. Berzins, J. I. Locus of control in dimensional versus typological perspectives. In H. L. Minton (Chair), *Beliefs in locus of control: Unidimensional or multidimensional?* Symposium presented at the meeting of the American Psychological Association, Montreal, August 1973.
7. Berzins, J. I., Welling, M. A., & Wetter, R. E. *Androgynous vs. traditional sex roles and the interpersonal behavior circle*. Paper presented at the meeting of the American Psychological Association, Washington, D.C., August-September 1976.

References

- Bakan, D. *The duality of human existence*. Chicago: Rand McNally, 1966.
- Bem, S. L. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1974, 42, 155-162.
- Bem, S. L. Sex-role adaptability: One consequence of psychological androgyny. *Journal of Personality and Social Psychology*, 1975, 31, 634-643.
- Bem, S. L. Probing the promise of androgyny. In A. G. Kaplan & J. P. Bean (Eds.), *Beyond sex-role stereotypes*. Boston: Little, Brown, 1976.
- Bem, S. L. On the utility of alternative procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1977, 45, 196-205.
- Bem, S. L., & Bem, D. J. Case study of a nonconscious ideology: Training the woman to know her place. In D. J. Bem (Ed.), *Beliefs, attitudes, and human affairs*. Belmont, Calif.: Brooks/Cole, 1970.
- Bem, S. L., & Lenney, E. Sex typing and the avoidance of cross-sex behavior. *Journal of Personality and Social Psychology*, 1976, 33, 48-54.
- Berzins, J. I. Therapist-patient matching. In A. S. Gurman & A. M. Razin (Eds.), *Effective psychotherapy*. New York: Pergamon Press, in press.
- Berzins, J. I., Barnes, D. F., Cohen, D. I., & Ross, W. F. A reappraisal of the A-B therapist "type" distinction in terms of the Personality Research Form. *Journal of Consulting and Clinical Psychology*, 1971, 36, 360-369.
- Block, J. H. Conceptions of sex role: Some cross-cultural and longitudinal perspectives. *American Psychologist*, 1973, 28, 512-527.
- Campbell, D. T. On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist*, 1975, 30, 1103-1126.
- Carlson, R. Sex differences in ego functioning: Exploratory studies of agency and communion. *Journal of Consulting and Clinical Psychology*, 1971, 37, 267-277.
- Carson, R. C. *Interaction concepts of personality*. Chicago: Aldine, 1969.
- Constantinople, A. Masculinity-femininity: An exception to a famous dictum? *Psychological Bulletin*, 1973, 80, 389-407.
- Cosentino, F., & Heilbrun, A. B. Anxiety correlates of sex-role identity in college students. *Psychological Reports*, 1964, 14, 729-730.
- Crowne, D. P., & Marlowe, D. *The approval motive*. New York: Wiley, 1964.
- Harman, H. H. *Modern factor analysis*. Chicago: University of Chicago Press, 1967.
- Heilbrun, A. B. Measurement of masculine and feminine sex role identities as independent dimensions. *Journal of Consulting and Clinical Psychology*, 1976, 44, 183-190.
- Heilbrun, A. B., & Thompson, N. L., Jr. Sex-role identity and male and female homosexuality. *Sex Roles*, 1977, 3, 65-79.
- Heist, P., & Yonge, G. *The Omnibus Personality Inventory: Form F*. New York: Psychological Corporation, 1968.
- Jackson, D. N. *Personality Research Form manual*. Goshen, N.Y.: Research Psychologists Press, 1967.
- Jung, C. G. *Two essays on analytical psychology*. New York: Meridian Books, 1956.
- Kaplan, A. G. Androgyny as a model of mental health for women: From theory to therapy. In A. G. Kaplan & J. P. Bean (Eds.), *Beyond sex-role stereotypes*. Boston: Little, Brown, 1976.
- Kaplan, A. G., & Bean, J. P. (Eds.) *Beyond sex-role stereotypes*. Boston: Little, Brown, 1976.
- Kelly, J. A., & Worell, L. Parent behaviors related to masculine, feminine, and androgynous sex role orientations. *Journal of Consulting and Clinical Psychology*, 1976, 44, 843-851.
- LaTorre, R. A. The psychological assessment of gender identity and gender role in schizophrenia. *Schizophrenia Bulletin*, 1976, 2, 266-285.
- Leary, T. *Interpersonal diagnosis of personality*. New York: Ronald Press, 1957.
- Parsons, T., & Bales, R. F. *Family, socialization, and interaction process*. New York: Free Press, 1955.
- Pinneau, S. R., & Newhouse, A. Measure of invariance and comparability in factor analysis for fixed variables. *Psychometrika*, 1964, 29, 271-281.
- Pleck, J. H. Masculinity-femininity: Current and alternative paradigms. *Sex Roles*, 1975, 1, 161-178.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Spence, J. T., Helmreich, R., & Stapp, J. Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 1975, 32, 29-39.
- Strahan, R. F. Remarks on Bem's measurement of psychological androgyny: Alternative methods and a supplementary analysis. *Journal of Consulting and Clinical Psychology*, 1975, 43, 568-571.
- Welling, M. A. *Measurement of psychological androgyny: Convergences between the PRF ANDRO scale and Bem's Sex-Role Inventory*. Unpublished master's thesis, University of Kentucky, 1976.
- Wiggins, J. S. *Personality and prediction: Principles of personality assessment*. Reading, Mass.: Addison-Wesley, 1973.

Sensation Seeking in England and America: Cross-cultural, Age, and Sex Comparisons

Marvin Zuckerman
University of Delaware

Sybil Eysenck and H. J. Eysenck
Institute of Psychiatry, University of London

This study compared the factor structure of the Sensation-Seeking Scale (SSS) in English and American samples, and a new form of the SSS, applicable to both samples, was constructed. Three of the four factors showed good cross-national and cross-sex reliability. English and American males did not differ on the total SSS score, but American females scored higher than English females. Males in both countries scored higher than females on the total SSS score and on the Thrill and Adventure-Seeking and Disinhibition subscales. Significant age declines occurred for both sexes, particularly on Thrill and Adventure Seeking and Disinhibition.

The Sensation-Seeking Scale (SSS) was developed in an attempt to provide an operational measure of the construct *optimal level of stimulation* (OLS). The construct is an old one, first formulated by Wundt (1873) to explain the curvilinear relationship between affective reactions and intensities of stimulation. After lying dormant for about 80 years, the OLS resurfaced in the 1950s and early 1960s in many theories, including those of Berlyne (1960), Fiske and Maddi (1961), Hebb (1955), Leuba (1955), Malmö (1959), McClelland, Atkinson, Clark, and Lowell (1953), and Schlosberg (1954). Berlyne, Fiske and Maddi, and Hebb, Malmö, and Schlosberg suggested that the idea of an optimal level of arousal (OLA) could be substituted for OLS, since the arousal construct could accommodate stimulus parameters such as novelty versus constancy, and complexity.

The first SSS (SSS II; Zuckerman, Kolin, Price, & Zoob, 1964) was developed with the idea of predicting responses to the experimen-

tal situation of sensory deprivation. It consisted of a General scale derived from factor analyses of many diverse kinds of items reflecting a positive reaction to or desire for stimulating, exciting, and novel kinds of experiences. This scale was rather heavily loaded with the risk-taking kind of items that subsequently became part of the Thrill and Adventure-Seeking subscale.

The SSS II was initially applied in sensory deprivation experiments, and the idea of stable individual differences in OLS and OLA became the central postulate in the theory developed by Zuckerman (1969) to explain differences in reactions to sensory deprivation. In the late 1960s, interest shifted from sensory deprivation to the broader construct-validity implications of the OLS-OLA measure (Zuckerman, Bone, Neary, Mangelsdorff, & Brustman, 1972; Zuckerman & Link, 1968). The SSS proved to have considerable validity for a variety of phenomena, ranging from design preferences to sexual and drug experiences and volunteering for unusual experiments or risky activities. Most of the research has been summarized in chapters by Zuckerman (1974, in press).

Studies by Farley (1967) and Zuckerman and Link (1968) suggested that there might be more than one factor of sensation seeking, and factor analyses were used in an attempt

The authors gratefully acknowledge the assistance of Judy Kasriel, who obtained the tests from the twins; Violet Beevers, who processed the data for the computer; and the subjects on both sides of the Atlantic who gave us their time and self-reports.

Requests for reprints should be sent to Marvin Zuckerman, Department of Psychology, University of Delaware, Newark, Delaware 19711.

to discover what these factors might be. Zuckerman (1971) reported the "dimensions of sensation seeking" found in samples of male and female undergraduates in the Philadelphia area. Four factors were identified, and three of them showed good reliability in their loadings across the sexes. The fourth factor, Boredom Susceptibility, was not clearly defined in the female group.

The first factor was called *Thrill and Adventure Seeking* (TAS), and it contained items expressing a desire to engage in sports or other activities involving speed or danger. The second factor was called *Experience Seeking* (ES), and it represented the seeking of experience through the mind and senses, travel, and a nonconforming life-style. The third factor was labeled *Disinhibition* (Dis), which seemed to represent the desire for social and sexual disinhibition as expressed in social drinking, partying, and variety in sexual partners. The fourth factor, called *Boredom Susceptibility* (BS), represented an aversion to repetition, routine, and dull people, and restlessness when things are unchanging. Examples of the sensation-seeking parts of the forced-choice items can be seen in Table 2.

Recently, Stewart and MacGriffith (1975) factor analyzed Form IV of the SSS, using 156 American undergraduates as subjects. The authors did not analyze male and female data separately. They found factors corresponding to TAS, ES, and BS, but Dis did not emerge as one of the first four factors. The authors noted that four factors only accounted for a quarter of the total variance and suggested that with no clear basis for only examining four factors, a general scale might suffice. Only four factors were analyzed in the study by Zuckerman (1971) because only four showed some reliable resemblance between the male and female samples. Stewart and MacGriffith suggested that different factors might exist in males and females. This is a possibility, of course, but Zuckerman was looking for factors that had broad applicability. Although the factor scales in Form IV remained substantially correlated, the General scale could not be used as an adequate substitute for the factor scales for two reasons: (a) Since it was based only on items that were contained in the Experimental Form I, it did not include

an adequate sampling of items from several factors based on new items written for Experimental Form III. (b) In the validity studies all of the scales, including the General scale, related to some of the validity criteria, but for other criteria, such as the physiological, only specific subscales showed significant relationships.

Before the Stewart and MacGriffith study, two factor studies were done on Form II of the SSS: one on an English sample (Farley, 1967) and one with a Japanese translation given to a Japanese sample (Ohkubo, Note 1). The general factor, defined by the unrotated factor loadings on the first factor, correlated .67 between English and American samples but only .35 between both American and English samples and the Japanese sample. A translated questionnaire cannot be expected to show very high cross-cultural factor reliability because of language and cultural differences. The reasonably high correlation between English and American samples was encouraging.

One of the aims of the present study was to examine the cross-cultural reliabilities of the factors in Form IV of the SSS, comparing the factor-analytic results from the original study (Zuckerman, 1971) with those from a large, socially heterogeneous sample of an English population. If the four factors originally derived from the American sample were meaningful and had a biological basis (Zuckerman, 1974), they would show cross-cultural as well as cross-sex stability.

A second purpose was to develop a shorter form of the SSS, based on the four factor analyses (English and American males and females). In this form, a total score, derived from the summation of the four factor scores, could be substituted for the General scale in Forms II and IV, which did not contain an adequate sample of items from the Dis and BS factors.

Assuming that the same factors could be found in English and American samples, the third question concerned cultural differences in scores on the sensation-seeking scales. Ohkubo (Note 1) found that both Japanese and Thai students (Thai data from Berkowitz, 1967) scored lower than American students on the General scale of Form II. These

differences must be interpreted cautiously, since they were based on translated sensation-seeking scales. Farley and Farley (1967) found that a sample of English male industrial apprentices and civil servants scored very close to the mean for American college males, despite the differences in socioeconomic and educational levels. The question remains: Do English and American subjects resemble each other on new subscales developed on the basis of factor similarity between the national samples?

Sex differences have been found on the General scale in American, Japanese, and Thai samples and on the factor scales in American college samples (Zuckerman, 1974, in press). A fourth purpose of the study was to see if similar sex differences would be found in the English sample.

A number of studies summarized by Zuckerman (1974, in press) have reported negative correlations between age and the sensation-seeking General scale when the samples covered a wide age range. However, none of these studies have studied age decline in a systematic fashion in all of the sensation-seeking scales. A fifth aim of this study was to examine age decline in the SSS subscales in the English sample. Adequate data are not yet available to do this kind of analysis in an American sample.

Method

Subjects

The English subjects consisted of 254 males and 693 females from the Maudsley Twin Register, ranging in age from 16 to 70. These subjects were used for the following reasons: (a) This group could provide data for a genetic analysis of the SSS; (b) they had previously taken the Eysenck Personality Questionnaire (EPQ) and thus could provide comparison data on these two instruments; (c) they covered a wide age range, thus providing our first good data for age comparison in a normal population; and (d) they were an interested and cooperative group, having answered previous questionnaires by mail. The SSS Form IV was sent out by mail to twins who had previously taken the EPQ. The return rate was about 80%.

The question can be raised as to whether or not twins are a special population. Previous studies showed that twins from the Maudsley Twin Register have normal patterns of scores on personality tests (Eysenck, 1976).

After the data from Form IV were analyzed in the English sample, a new form (Form V) of 40 items was constructed. This form, along with the EPQ, was given to 97 male and female undergraduates from two large sections of undergraduate psychology at the University of Delaware. Most of these students fell in the age range of 17-25, and the modal range was 18-20. The SSS was given first, and the EPQ was given second.

The 72-item SSS Form IV item responses were intercorrelated and factor analyzed in the English male and female samples separately. They were rotated obliquely to simple structure using the promax method, and the loadings were compared with the loadings from the previous study by Zuckerman (1971). Product-moment correlations, the principal components method, and oblique rotations were used in both studies. The previous study (Zuckerman, 1971) factored the SSS Experimental Form III, which contained 113 items, whereas the current study factored Form IV, containing 72 items. The factor structure might have shown some differences due to the extra 41 items in the previous study. These 41 items were not included in Form IV because they showed no loadings of any size on the four primary factors. This means that they contributed little to the interpreted factor structure, so the difference between the two forms should not have introduced much distortion in the comparison of factors across samples.

The sample on which the American factor analysis was based consisted of 160 male and 172 female undergraduates from psychology courses at Temple University. The English sample was much more heterogeneous as to age and education, and differences in the factor structure could have occurred because of the difference in populations. But to repeat the argument used in regard to sex differences in factor structure: If factors are meaningful and have some biological basis, they should be generalizable to a broad range of the population.

Our design allowed for sex and national comparisons of factor reliability: English males with American males, English females with American females, English males with English females, American males with American females, English males with American females, and English females with American males.

Results

Factor Reliabilities

The rotated factor loadings on the first four factors in the six samples were correlated over 71 items. One item from the General scale was missing in the 113-item Form III given to the American groups, but this was not an item contained in any of the factor scales. The matrix of correlations was 16×16 (4 samples \times 4 factors).

The factor reliability coefficients from this matrix are shown in Table 1, which also gives

Table 1
Factor Reliability Coefficients

Comparisons between	Sensation-Seeking Scale factors			
	TAS	ES	Dis	BS
Eng M/Eng F	.90	.68	.79	.70
Eng M/U.S. M	.87	.72	.60	.37
Eng M/U.S. F	.67	.51	.66	.51
Eng F/U.S. M	.88	.68	.66	.65
Eng F/U.S. F	.72	.75	.77	.30
U.S. M/U.S. F	.73	.67	.79	.01
Range of cross-factor correlations	.06 to -.55	.09 to -.49	.41 to -.49	.44 to -.55

Note. Eng = English, U.S. = American, M = males, F = females; TAS = Thrill and Adventure Seeking, ES = Experience Seeking, Dis = Disinhibition, BS = Boredom Susceptibility.

the range of other correlations between different factors within a sample and between different factors across samples. As in the multimethod-multitrait validity model (Campbell & Fiske, 1959), the correlations between the same factors across samples should be significant and exceed the correlations between different factors (traits in the multimethod-multitrait model). Of course, reliability coefficients should also be high as well as being significant.

Three of the four factors clearly met these criteria of factor reliability. Thrill and Adventure Seeking, Experience Seeking, and Disinhibition all showed significant and reasonably high resemblance between the four national and sex samples. With the exception of only 1 of the 18 correlations for these three scales, the coefficients all exceeded .60 and all fell above the range of the cross-factor correlations.

The case was not as clear for the Boredom Susceptibility factor. Although the factor was fairly similar in English males and females, it showed no correspondence in American males and females, and two of the four cross-national comparisons showed weak ($< .40$), if significant, resemblance. Only three of the coefficients exceeded the range of the cross-factor correlations for this factor.

Construction of Form V

On the basis of the four factor analyses, an attempt was made to select items for a new

form (Form V), with the aim of using 10 items for each factor that met the criteria of having a primary loading on the same factor in all samples and loadings exceeding .30 in magnitude. Such scales would have the greatest value for the cross-cultural comparisons.

There was no problem in meeting these criteria for TAS, in which all but one of the items (21) had loadings over .30 on the factor in all four samples. When a choice had to be made between items having similarly high factor loadings, an attempt was made to diversify the content in the new scales.

For Dis, one item (47) had to be selected that did not meet the criteria in the American male sample and another (66) that did not in the English male sample. The BS scale had previously been defined by male loadings, since the factor was not well defined in the American females. Another item (12) did not meet the criteria in the English males. One item (61) had to be included even though it did not meet the criteria in the American analyses. The most radical change was on the ES scale, on which three new items (11, 15, and 25) had to be included even though they did not meet the criteria in all the samples. In some cases items with loadings of slightly less than .30 had to be used, but they always had their highest loading on the relevant factor.

In this manner, 10 items were selected from each of the four primary factors in the SSS to comprise Form V. A total score for this new form can be obtained by summing the

four subscale scores, but Form V contains no general scale as did Form IV.

Table 2 shows the sensation-seeking choice on each item selected, with the factor loadings of that item on the appropriate factor in each of the four samples. In comparing these loadings it should be remembered that the factor analyses were of 72 items in the English sample and 113 in the American sample, so that some of the total variance may have been on items not directly compared across samples.

Scale Reliabilities on Forms IV and V

Table 3 shows internal reliabilities in the English and American Temple University samples for Form IV and the English and American University of Delaware samples for Form V. The reliabilities of the English and American samples were quite similar for Form IV even though the structure of these scales was determined only by the American factor analyses.

The reliabilities of the Form V factor scales were expected to be somewhat lower because the scales were shorter, that is, 10-item scales as opposed to 14- and 18-item scales in Form IV. Actually the only substantial drop in reliability was on the ES scale where reliabilities fell from .7 and .8 to .6.

The most homogeneous scales, TAS and Dis, showed little loss of reliability in the new form; the ES reliability fell somewhat but was still within acceptable limits; the BS reliability remained at the borderline range of high .5, where it had been for American females in Form IV.

Correlations Between Scales

The correlations between the factor scales in Forms IV and V are shown in Table 4. These scale correlations had been rather high in Form IV, and it was hoped that the scales in Form V would have more independence, although some significant correlation was still expected.

Table 4 shows that the correlations among subscales, particularly among ES, Dis, and BS, that were quite high in Form IV were reduced in Form V. TAS continued to correlate with ES in Form IV but showed very low and

sometimes insignificant correlations with Dis and BS.

National and Sex Differences

Table 5 shows the comparisons of English and American male and female samples on Form V. Form V was felt to be most appropriate for these comparisons, since these factor scales were based on cross-national similarity of factors. Only the 16- to 19-year-old English subjects were used in these comparisons, since the American college students were mostly within this age range. Although the young English and American males were not different on the total SSS score, the American males were significantly higher on Experience Seeking, and the English males scored higher on Boredom Susceptibility. The American females were significantly higher than the English on the total score, the Thrill and Adventure-Seeking scale, and the Experience-Seeking scale.

Looking at sex differences within the two national groups, both English and American males were significantly higher than the females on the total score and on the TAS and Dis factor scores. The English males were higher than the English females on BS. There were no significant sex differences on the ES scale in either country.

Age Comparisons

Table 6 shows the mean scores of the males and females on the SSS, Form V within each age group of the English sample. Figure 1

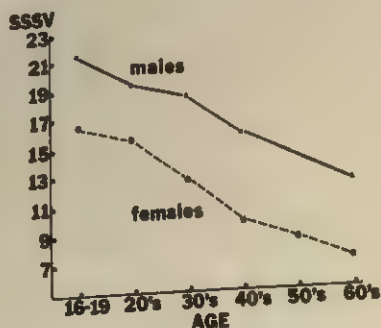


Figure 1. Changes in sensation-seeking total scores as a function of age. (SSSV = Sensation-Seeking Scale Form V.)

Table 2

Loadings of Items Selected for Form V of the Sensation-Seeking Scale

No. on Form IV	Item	Loading			
		English		American	
		M	F	M	F
Thrill and Adventure Seeking					
10	I often wish I could be a mountain climber.	.68	.47	.44	.31
21	I sometimes like to do things that are a little frightening.	.43	.36	.50	.25
29	I would like to take up the sport of water skiing.	.62	.73	.71	.64
31	I would like to try surfboard riding.	.64	.64	.76	.56
35	I would like to learn to fly an airplane.	.45	.64	.63	.33
37	I would like to go scuba diving.	.59	.58	.64	.36
43	I would like to try parachute jumping.	.63	.69	.66	.32
53	I like to dive off the high board.	.53	.54	.54	.45
69	I would like to sail a long distance in a small but sea-worthy sailing craft.	.52	.46	.54	.38
71	I think I would enjoy the sensations of skiing very fast down a high mountain slope.	.65	.58	.57	.52
Experience Seeking					
11	I like some of the earthy body smells.	.40	.39	.46	.16
15	I like to explore a strange city or section of town myself, even if it means getting lost.	.28	.56	.23	.48
18	I have tried marijuana or would like to.	.51	.47	.60	.58
19	I would like to try some of the new drugs that produce hallucinations.	.34	.36	.60	.44
25	I like to try new foods that I have never tasted before.	.60	.33	.23	.34
33	I would like to take off on a trip with no preplanned or definite routes or timetables.	.37	.39	.40	.41
34	I would like to make friends in some of the "far-out" groups like artists or "hippies."	.57	.35	.48	.45
38	I would like to meet some persons who are homosexual (men or women).	.56	.47	.56	.54
51	I often find beauty in the "clashing" colors and irregular form of modern painting.	.53	.36	.50	.45
68	People should dress in individual ways even if the effects are sometimes strange.	.42	.41	.47	.49
Disinhibition					
6	I like wild "uninhibited" parties.	.49	.57	.53	.54
22	I enjoy the company of real "swingers."	.50	.41	.38	.50
23	I often like to get high (drinking liquor or smoking marijuana).	.39	.59	.41	.46
47	I like to have new and exciting experiences and sensations even if they are a little unconventional or illegal.	.41	.31	.12	.32
54	I like to date members of the opposite sex who are physically exciting.	.46	.58	.45	.32
55	Keeping the drinks full is the key to a good party.	.45	.54	.35	.49
59	A person should have considerable sexual experience before marriage.	.35	.43	.42	.32
60	I could conceive of myself seeking pleasures around the world with the "jet set."	.46	.37	.40	.42
64	I enjoy watching many of the "sexy" scenes in movies.	.54	.68	.39	.54
66	I feel best after taking a couple of drinks.	.26	.47	.42	.33

Table 2 (continued)

No. on Form IV		Item	Loading			
			English		American	
			M	F	M	F
Boredom Susceptibility						
7	I can't stand watching a movie that I've seen before.	.47	.34	.35	.22	
12	I get bored seeing the same old faces.	.13	.50	.50	.12	
16	When you can predict almost everything a person will do and say, he or she must be a bore.	.45	.47	.40	.26	
17	I usually don't enjoy a movie or a play where I can predict what will happen in advance.	.42	.41	.56	.03	
27	Looking at someone's home movies or travel slides bores me tremendously.	.65	.35	.30	.24	
46	I prefer friends who are excitingly unpredictable.	.39	.34	.33	.15	
52	I get very restless if I have to stay around home for any length of time.	.39	.39	.31	.04	
57	The worst social sin is to be a bore.	.61	.24	.34	.12	
61	I like people who are sharp and witty even if they do sometimes insult others.	.47	.29	.25	.05	
70	I have no patience with dull or boring persons.	.54	.37	.32	.39	

Note. M = male, F = female. Only the sensation-seeking choices in the forced-choice items are presented here; a copy of Form V can be obtained from the first author.

shows the data for the total score, and Figure 2 shows the trends for the separate scales.

The *F* values from simple analyses of variance between the age groups for each sex are shown in Table 6. The age differences in sensation seeking were significant for the total score and on all scales for the females. The age change was significant for the total score and the Thrill and Adventure-Seeking and Disinhibition scales for males but was not significant for Experience Seeking and Boredom Susceptibility. Examining Figure 2, it is apparent that the fall in SSS scores was more pronounced for the TAS and Dis scales than for the ES and BS scales.

Discussion

Considering the differences in populations sampled in England and America and the additional items used in Form III on the American sample, the amount of cross-national and cross-sex correspondence in the SSS factors is impressive. Even the Boredom Susceptibility factor, which had not shown cross-sex reliability in the American sample, did show such reliability in the English one. The data argue strongly for the meaningfulness of the factor

scales of the SSS Form IV. This is not to say that we can generalize to other cultures. The status of the factors in translated forms of the SSS is still an open question.

On the basis of the factor stabilities, we were able to construct a new shorter Form V of the SSS, with a total score balanced for the four factors. This new form has the advantage of reducing the interscale correlations between component factor scores with little loss in reliability of these scores. Form V should prove useful in further research on sensation seeking on both sides of the Atlantic.

Although there were educational differences between the younger American and English samples that were compared, education has not proved to be a highly significant factor in the SSS (Farley & Farley, 1967; Kish & Busse, 1968). The English and American males did not differ on the total score on Form V, but the pattern on the subscales was different, with Americans scoring higher on Experience Seeking and the English scoring higher on Boredom Susceptibility. Experience seeking seems to represent a style of life common in the 1960s that is still an influence in America in the 1970s but it is apparently not as important in England. The ES scale was

Table 3

Scale Reliabilities on Sensation-Seeking Scale Forms IV and V

Scale	Form IV				Form V			
	English ^a		American ^b		English ^a		American ^c	
	M	F	M	F	M	F	M	F
General	.72	.80	.75	.81	—	—	—	—
TAS	.83	.84	.85	.85	.81	.82	.77	.77
ES	.76	.78	.84	.88	.65	.67	.61	.61
Dis	.77	.75	.71	.75	.78	.77	.74	.76
BS	.62	.66	.75	.58	.65	.59	.57	.56
Total scores (Form V)	—	—	—	—	.83	.86	.84	.85

Note. M = males, F = females; TAS = Thrill and Adventure Seeking, ES = Experience Seeking, Dis = Disinhibition, BS = Boredom Susceptibility. All coefficients in the table are significant, $p < .01$.

^a Alpha coefficients were used; $ns = 254$ males and 693 females.

^b Split-half corrected coefficients were used; $ns = 160$ males and 170 females.

^c Internal consistency coefficients (from interitem rs) were used; $ns = 97$ males and 122 females.

the only one that did not show any hereditary influence in the preliminary study by Buchsbaum as reported in Zuckerman (1974). Experience seeking may have been most influenced by the educational differences between the two samples.

Disinhibition seemed to be less influenced by cultural differences than did other scales. Some unpublished data on racial differences in America (Kurtz & Zuckerman, Note 2) have shown that blacks are lower than whites on

TAS and BS but not on Dis or ES. Dis seems to be the most culture-free scale, and it is the one most highly related to certain psychophysiological variables (Zuckerman, 1974).

In contrast with the males, the American females were significantly higher than the English females on the total score and on the ES and TAS subscales. As with the males, no national differences were found on the Dis scale.

The new Form V shows more selective sex

Table 4

Correlations Between Subscales on Forms IV and V

Sensation-seeking scales correlated	Form IV				Form V			
	English ^a		American ^b		English ^a		American ^c	
	M	F	M	F	M	F	M	F
TAS × ES	.42	.52	.39	.37	.27	.42	.27	.39
TAS × Dis	.22	.35	.35	.21	.25	.35	.14*	.29
TAS × BS	.28	.36	.25	.28	.10*	.20	.06*	.18
ES × Dis	.54	.57	.54	.51	.32	.47	.24**	.40
ES × BS	.57	.59	.51	.62	.21	.29	.26	.37
Dis × BS	.45	.50	.44	.34	.42	.48	.37	.40

Note. M = males, F = females; TAS = Thrill and Adventure Seeking, ES = Experience Seeking, Dis = Disinhibition, BS = Boredom Susceptibility. All correlations were significant ($p < .01$), unless otherwise indicated.

^a $Ns = 254$ males and 693 females.

^b $Ns = 160$ males and 170 females.

^c $Ns = 97$ males and 122 females.

* ns .

** $p < .05$.

Table 5
Comparisons of American College Sample with Younger English Twin Samples (Ages 16-19)
on Sensation-Seeking Scale Form V

Scale	Males					Females					Sex differences	
	Eng ^a		U.S. ^b		Eng vs. U.S. <i>t</i>	Eng ^c		U.S. ^d		Eng vs. U.S. <i>t</i>	Eng	U.S.
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		<i>t</i>	<i>t</i>
TAS	7.4	2.6	7.8	2.3	1.40	5.6	3.0	6.9	2.6	3.45**	4.20**	2.73**
ES	4.1	2.1	4.8	2.0	3.03**	4.1	2.2	5.0	2.1	3.08**	.12	.73
Dis	6.2	3.1	5.9	2.6	.86	4.1	2.7	4.7	2.7	1.71	4.59**	3.21**
BS	3.8	2.5	3.2	2.0	2.76**	2.8	2.1	3.0	1.9	.64	2.91**	.65
Total	21.5	6.7	21.6	5.7	.17	16.6	7.2	19.6	6.6	3.22**	4.58**	2.44*

Note. Eng = English, U.S. = American; TAS = Thrill and Adventure Seeking, ES = Experience Seeking, Dis = Disinhibition, BS = Boredom Susceptibility.

^a *n* = 72.

^b *n* = 97.

^c *n* = 106.

^d *n* = 122.

* *p* < .05.

** *p* < .01.

differences than does the old Form IV, on which males were higher than females on all of the scales (Zuckerman, 1974). On Form V the replicated (across nations) sex differences were limited to the TAS and Dis scales. Males also scored significantly higher on the total score in both countries. The Dis scale showed the largest sex difference, even on Form IV. The difference on Dis can, of course, be interpreted as reflecting different kinds of socialization experiences of males and females in

both countries. However, the finding of a relation between gonadal hormones and Dis in a sample of American males (Daitzman, Zuckerman, Sammelwitz, & Venkateshu, Note 3) suggests that biological factors may also play a role in differences on this personality dimension.

The decline in sensation seeking with age was predicted in the theory formulated by Zuckerman (1969). In the chapter by Zuckerman (1974), it was predicted that TAS and

Table 6
Mean Scores of English Males and Females by Age Groups

Ages	<i>n</i>		Total score		TAS		ES		Dis		BS	
	<i>M</i>	<i>F</i>	<i>M</i>	<i>F</i>	<i>M</i>	<i>F</i>	<i>M</i>	<i>F</i>	<i>M</i>	<i>F</i>	<i>M</i>	<i>F</i>
16-19	72	106	21.5	16.6	7.4	5.6	4.1	4.1	6.2	4.1	3.8	5.1
20-29	119	250	19.3	15.4	6.6	4.4	4.4	4.2	4.9	4.0	3.5	2.8
30-39	25	145	18.4	12.3	5.7	3.4	4.5	4.0	4.6	2.9	3.6	2.1
40-49	26	89	15.8	10.7	4.3	2.6	4.0	3.7	4.6	2.1	3.0	2.3
50-59	—	69	—	8.5	—	2.3	—	2.8	—	1.4	—	2.0
60+	12	34	12.4	7.0	3.2	1.7	2.7	2.2	3.3	1.0	3.3	2.0
<i>F</i> between age groups ^a												
	9.3*	36.0*	11.6*	40.3*	2.0	11.6*	4.6*	32.5*	<1	5.1*		

Note. *M* = males, *F* = females; TAS = Thrill and Adventure Seeking, ES = Experience Seeking, Dis = Disinhibition, BS = Boredom Susceptibility.

^a *df*: males = 4/249, females = 4/688.

* *p* < .001.

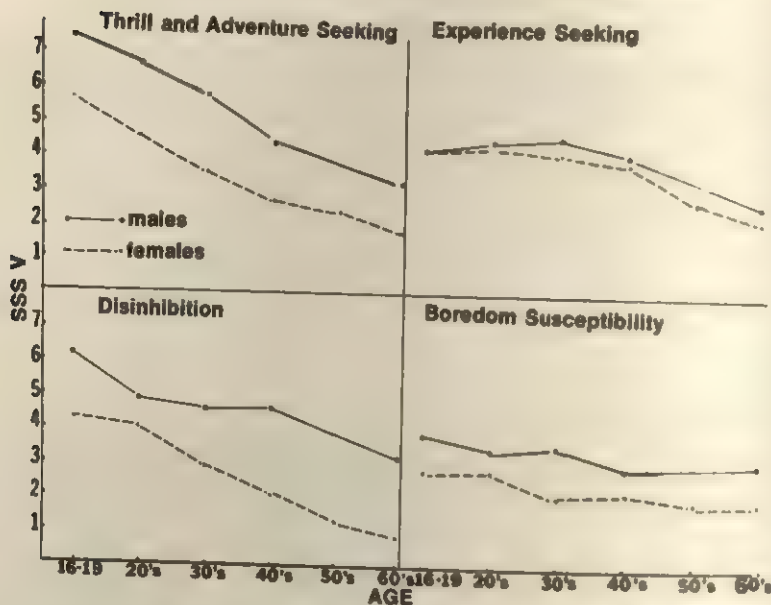


Figure 2. Changes in sensation-seeking subscale scores as a function of age. (SSSV = Sensation-Seeking Scale Form V.)

ES would decline at faster rates than Dis and BS. The results showed a greater decline for TAS and Dis than for ES and BS. A regression analysis showed that age accounted for 18% of the variance for Dis, 21% for TAS, 5% for ES, and 3% for BS. There was a clear linear decrease in sensation seeking with age on the total SSS score.

What is the basis of the increasing cautiousness and conservatism of age? It might simply reflect the mellowing effect of accumulated experience. But many biological changes also occur with age, including a slowing of cortical activity and diminution of gonadal hormone output. Eysenck (1967) has suggested that biological factors may affect attitudinal traits. It can be hypothesized that the same biological factors that are prominent in aging affect the sensation-seeking tendency.

The other part of the age postulate (Zuckerman, 1969) suggests that sensation seeking increases from childhood to adolescence. Farley and Cox (1971) found no increase from ages 14 to 17. A new instrument will have to be developed for younger children in order to test this hypothesis.

Reference Notes

1. Ohkubo, Y. Application of the sensation-seeking scale to Japanese students. Paper presented at the

36th annual convention of the Japanese Psychological Association, Osaka, Japan, August 1972.

2. Kurtz, J., & Zuckerman, M. Racial and sex differences in sensation seeking in a college population. Manuscript submitted for publication, 1977.
3. Daitzman, R. J., Zuckerman, M., Sammelwitz, P., & Venkateshu, G. Gonadal hormones and sensation-seeking: Further evidence for a biological basis of the sensation-seeking motive. Paper presented at the meeting of the Eastern Psychological Association, New York, April 1976.

References

- Berkowitz, W. R. Use of the Sensation-Seeking Scale with Thai subjects. *Psychological Reports*, 1967, 20, 635-641.
- Berlyne, D. E. *Conflict, arousal, and curiosity*. New York: McGraw-Hill, 1960.
- Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Eysenck, H. J. *The biological basis of personality*. Springfield, Ill.: Charles C Thomas, 1967.
- Eysenck, H. J. *Sex and personality*. Austin: University of Texas Press, 1976.
- Farley, F. H. Social desirability and dimensionality in the Sensation-Seeking Scale. *Acta Psychologica*, 1967, 26, 89-96.
- Farley, F. H., & Cox, S. O. Stimulus-seeking motivation in adolescents as a function of age and sex. *Adolescence*, 1971, 6, 207-218.
- Farley, F. H., & Farley, S. V. Extroversion and stimulus-seeking motivation. *Journal of Consulting Psychology*, 1967, 31, 215-216.

- Fiske, D. W., & Maddi, S. R. *Functions of varied experience*. Homewood, Ill.: Dorsey Press, 1961.
- Hebb, D. O. Drives and the CNS (conceptual nervous system). *Psychological Review*, 1955, 62, 243-254.
- Kish, G. B., & Busse, W. Correlates of stimulus seeking: Age, education, intelligence, and aptitudes. *Journal of Consulting and Clinical Psychology*, 1968, 32, 633-637.
- Leuba, C. Toward some integration of learning theories: The concept of optimal stimulation. *Psychological Reports*, 1955, 1, 27-33.
- Malmö, R. B. Activation: A neurophysiological dimension. *Psychological Review*, 1959, 66, 367-386.
- McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. *The achievement motive*. New York: Appleton-Century-Crofts, 1953.
- Schlossberg, H. Three dimensions of emotion. *Psychological Review*, 1954, 61, 81-88.
- Stewart, D. W., & MacGriffith, G. Factor analysis of Zuckerman's Sensation-Seeking Scale. *Psychological Reports*, 1975, 37, 849-850.
- Wundt, W. M. *Grundzuge der physiologischen psychologie*. Leipzig, German Democratic Republic: Engelman, 1873.
- Zuckerman, M. Theoretical formulations. In J. P. Zubek (Ed.), *Sensory deprivation: 15 years of research*. New York: Appleton-Century-Crofts, 1969.
- Zuckerman, M. Dimensions of sensation seeking. *Journal of Consulting and Clinical Psychology*, 1971, 36, 45-52.
- Zuckerman, M. The sensation-seeking motive. In B. Maher (Ed.), *Progress in experimental personality research* (Vol. 7). New York: Academic Press, 1974.
- Zuckerman, M. Sensation seeking. In H. London & J. Exner (Eds.), *Dimensions of personality*. New York: Wiley, in press.
- Zuckerman, M., Bone, R. N., Neary, R., Mangelsdorff, D., & Brustman, B. What is the sensation seeker? Personality trait and experience correlates of the Sensation-Seeking scales. *Journal of Consulting and Clinical Psychology*, 1972, 39, 308-321.
- Zuckerman, M., Kolin, E. A., Price, L., & Zoob, I. Development of a Sensation-Seeking Scale. *Journal of Consulting Psychology*, 1964, 28, 477-482.
- Zuckerman, M., & Link, K. Construct validity for the Sensation-Seeking Scale. *Journal of Consulting and Clinical Psychology*, 1968, 32, 420-426.

Received April 20, 1977 ■

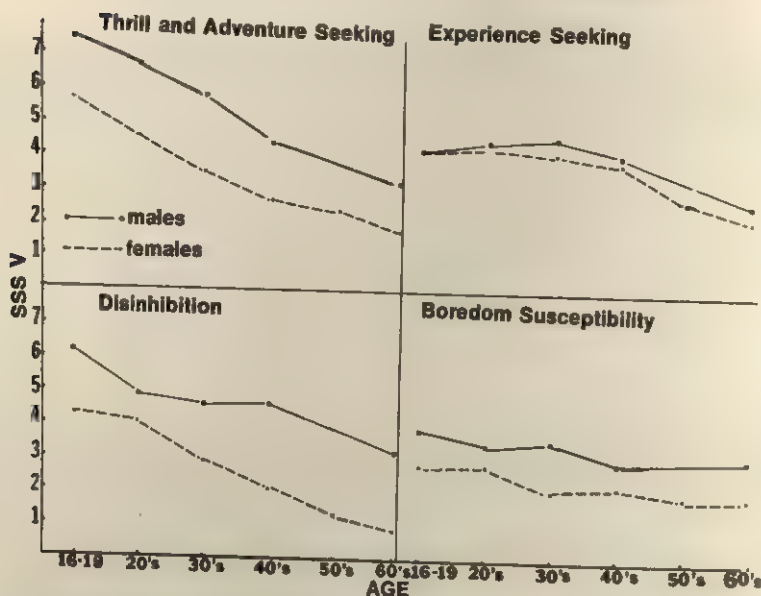


Figure 2. Changes in sensation-seeking subscale scores as a function of age. (SSSV = Sensation-Seeking Scale Form V.)

ES would decline at faster rates than Dis and BS. The results showed a greater decline for TAS and Dis than for ES and BS. A regression analysis showed that age accounted for 18% of the variance for Dis, 21% for TAS, 5% for ES, and 3% for BS. There was a clear linear decrease in sensation seeking with age on the total SSS score.

What is the basis of the increasing cautiousness and conservatism of age? It might simply reflect the mellowing effect of accumulated experience. But many biological changes also occur with age, including a slowing of cortical activity and diminution of gonadal hormone output. Eysenck (1967) has suggested that biological factors may affect attitudinal traits. It can be hypothesized that the same biological factors that are prominent in aging affect the sensation-seeking tendency.

The other part of the age postulate (Zuckerman, 1969) suggests that sensation seeking increases from childhood to adolescence. Farley and Cox (1971) found no increase from ages 14 to 17. A new instrument will have to be developed for younger children in order to test this hypothesis.

Reference Notes

1. Ohkubo, Y. *Application of the sensation-seeking scale to Japanese students*. Paper presented at the

36th annual convention of the Japanese Psychological Association, Osaka, Japan, August 1972.

2. Kurtz, J., & Zuckerman, M. *Racial and sex differences in sensation seeking in a college population*. Manuscript submitted for publication, 1977.
3. Daitzman, R. J., Zuckerman, M., Sammelwitz, P., & Venkateshu, G. *Gonadal hormones and sensation-seeking: Further evidence for a biological basis of the sensation-seeking motive*. Paper presented at the meeting of the Eastern Psychological Association, New York, April 1976.

References

- Berkowitz, W. R. Use of the Sensation-Seeking Scale with Thai subjects. *Psychological Reports*, 1967, 20, 635-641.
- Berlyne, D. E. *Conflict, arousal, and curiosity*. New York: McGraw-Hill, 1960.
- Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Eysenck, H. J. *The biological basis of personality*. Springfield, Ill.: Charles C Thomas, 1967.
- Eysenck, H. J. *Sex and personality*. Austin: University of Texas Press, 1976.
- Farley, F. H. Social desirability and dimensionality in the Sensation-Seeking Scale. *Acta Psychologica*, 1967, 26, 89-96.
- Farley, F. H., & Cox, S. O. Stimulus-seeking motivation in adolescents as a function of age and sex. *Adolescence*, 1971, 6, 207-218.
- Farley, F. H., & Farley, S. V. Extroversion and stimulus-seeking motivation. *Journal of Consulting Psychology*, 1967, 31, 215-216.

- Fiske, D. W., & Maddi, S. R. *Functions of varied experience*. Homewood, Ill.: Dorsey Press, 1961.
- Hebb, D. O. Drives and the CNS (conceptual nervous system). *Psychological Review*, 1955, 62, 243-254.
- Kish, G. B., & Busse, W. Correlates of stimulus seeking: Age, education, intelligence, and aptitudes. *Journal of Consulting and Clinical Psychology*, 1968, 32, 633-637.
- Leuba, C. Toward some integration of learning theories: The concept of optimal stimulation. *Psychological Reports*, 1955, 1, 27-33.
- Malmö, R. B. Activation: A neurophysiological dimension. *Psychological Review*, 1959, 66, 367-386.
- McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. *The achievement motive*. New York: Appleton-Century-Crofts, 1953.
- Schlosberg, H. Three dimensions of emotion. *Psychological Review*, 1954, 61, 81-88.
- Stewart, D. W., & MacGriffith, G. Factor analysis of Zuckerman's Sensation-Seeking Scale. *Psychological Reports*, 1975, 37, 849-850.
- Wundt, W. M. *Grundzüge der physiologischen psychologie*. Leipzig, German Democratic Republic: Engelman, 1873.
- Zuckerman, M. Theoretical formulations. In J. P. Zubek (Ed.), *Sensory deprivation: 15 years of research*. New York: Appleton-Century-Crofts, 1969.
- Zuckerman, M. Dimensions of sensation seeking. *Journal of Consulting and Clinical Psychology*, 1971, 36, 45-52.
- Zuckerman, M. The sensation-seeking motive. In B. Maher (Ed.), *Progress in experimental personality research* (Vol. 7). New York: Academic Press, 1974.
- Zuckerman, M. Sensation seeking. In H. London & J. Exner (Eds.), *Dimensions of personality*. New York: Wiley, in press.
- Zuckerman, M., Bone, R. N., Neary, R., Mangelsdorff, D., & Brustman, B. What is the sensation seeker? Personality trait and experience correlates of the Sensation-Seeking scales. *Journal of Consulting and Clinical Psychology*, 1972, 39, 308-321.
- Zuckerman, M., Kolin, E. A., Price, L., & Zoob, I. Development of a Sensation-Seeking Scale. *Journal of Consulting Psychology*, 1964, 28, 477-482.
- Zuckerman, M., & Link, K. Construct validity for the Sensation-Seeking Scale. *Journal of Consulting and Clinical Psychology*, 1968, 32, 420-426.

Received April 20, 1977 ■

Validation of the Beck Depression Inventory in a University Population Using Psychiatric Estimate as the Criterion

William Bumberry and J. M. Oliver
Saint Louis University

James N. McClure
Washington University

This study investigated the utility of the Beck Depression Inventory for survey use in a college population by determining its concurrent validity, using psychiatric rating of depth of depression as the criterion ($N = 56$). Interrater reliability of psychiatric estimate as measured by a Pearson product-moment correlation coefficient was .62, perhaps because the primary and secondary raters used different diagnostic procedures. The Pearson product-moment correlation coefficient between the inventory and the psychiatric rating was .77. These findings indicate that the Beck Depression Inventory is indeed a valid instrument for use in a college population. The Pearson product-moment correlation coefficient between the inventory and the psychiatric estimate fell to .30 in a second sample in which 1-14 days intervened between administration of the inventory and the psychiatric interview ($n = 27$). This attenuation in subjects who experienced a time delay is consistent with the nature of the depression inventory as a measure of state as opposed to trait depression. The apparent decline in measured depression additionally suggests the need for longitudinal study to determine its course and outcome.

Depression is widely viewed as the most frequently occurring psychic disorder among college students. Seligman (1973) contends that it is not only the most common of the psychological dysfunctions among students, but it is also increasing in frequency. Very little information is available concerning the prevalence of depression in college populations.

Recently, however, Oliver, Croghan, and Katz (Note 1) have estimated the prevalence of depression in college students by administering the Beck Depression Inventory (Beck, 1970; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) to a random sample of freshmen and sophomores at four private, medium-

sized, coeducational, urban universities. When the criterion for depression was set at a number or intensity of symptoms associated with a diagnosis of depression in a psychiatric population, 23% of the respondents qualified as at least mildly depressed. These findings suggest that the rate of depression may be as much as 50% higher in college students than in American adults between the ages of 18 and 74, in whom the prevalence of depression was cited as 15% in a special report on depression issued by the National Institute of Mental Health (1973). Thus, there is a large discrepancy between the estimates of the rate prevailing in the unselected population reported by the National Institute of Mental Health and that found by Oliver et al. using the Beck Depression Inventory.

The Beck Depression Inventory is a clinically derived self-report inventory of depression designed for use in psychiatric populations. It was constructed to assess the current depth of depression, whether or not depres-

The completion of this study was made possible by the generous participation of James Halikas and Amos Weiner, both of Washington University Medical School.

Requests for reprints should be sent to William Bumberry, who is now at 1970 Latham, Apartment 44, Mountain View, California 94040.

Copyright 1978 by the American Psychological Association, Inc. All rights of reproduction in any form reserved.

sion is viewed as the primary diagnosis, and consists of 21 items covering affective, cognitive, motivational, and physiological areas of depressive symptomatology. The range of possible scores extends from 0 to 63, with scores of 0-9 being categorized by Beck as not depressed, 10-15 as mildly depressed, 16-23 as moderately depressed, and 24-63 as severely depressed.

Two validation studies were reported by Beck et al. (1961); they were conducted in psychiatric populations, with psychiatric assessment of depth of depression comprising the criterion measure with which the depression inventory was compared. Biserial correlation coefficients of .65 and .67 were obtained ($ns = 226$ and 183 , respectively). These validation studies may be largely responsible for the rather widespread application that the depression inventory has recently enjoyed, having already been used as a criterion measure in well over 100 published studies (Beck & Beck, 1972). In spite of its popularity, however, it should be noted that the depression inventory was designed for use with psychiatric populations and has not been demonstrated to be a valid measure of depression for survey use in a college or any other unselected population.

The purpose of this study was to assess the applicability of the Beck Depression Inventory for survey use in this population. The method consisted of determination of the concurrent or diagnostic validity of the depression inventory, using psychiatric rating of depth of depression as the criterion measure.

Method

Subjects

Subjects in this study consisted of 56 university students (27 males and 29 females) at both undergraduate and graduate levels from two medium-sized, private, coeducational, urban universities (Washington University and Saint Louis University) located in the same midwestern city. The Scholastic Aptitude Test scores at these two institutions averaged 550 Verbal, 590 Math and 500 Verbal, 500 Math, respectively.

An attempt was made (although unsuccessfully) to include an equal number of subjects from each of the four categories of depression as designated by Beck. Since 77% and 16% of randomly selected

lowerclassmen can be expected to score in Beck's normal and mild categories, respectively (Oliver et al., Note 1), subjects in these two categories ($ns = 16$ and 19 , respectively) were obtained through a random sampling procedure at Saint Louis University. The university phone directory, constituting the registrar's list, provided the pool from which subjects were randomly selected. Each subject from this pool was personally contacted.

Since only 6% and 1% of randomly sampled lower classmen can be expected to fall in the moderately and severely depressed categories, respectively, this strategy had to be supplemented to procure a sufficient number of subjects in these categories ($ns = 17$ and 4 , respectively). More extensive sampling was obtained by administering the depression inventory to various large undergraduate and graduate classes. Finally, with the cooperation of the psychiatric division of the Samuel B. Grant Health Service at Washington University, students referred for psychiatric evaluation with and without depression were referred to the study. These subjects were administered both the depression inventory and the psychiatric interview for depth of depression before any other diagnostic or therapeutic procedures were introduced.

Materials

The Beck Depression Inventory as originally published (Beck et al., 1961) was used. After the subjects had completed the entire inventory, the duration of each symptom was also obtained.

The structured interview used by the primary psychiatrist was developed and has been used for over 2 decades by the Washington University Department of Psychiatry. Interrater reliability of information derived from this interview has been discussed by several investigators (Feighner, Robins, Guze, Woodruff, Winokur, & Munoz, 1972; Helzer et al. 1977). Data obtained from this interview form the basis of the forthcoming revision of the *APA Diagnostic and Statistical Manual* (APA, in press) and of numerous publications concerning a wide range of disorders from anxiety neurosis to organic brain syndrome and including depression.

Those psychiatrists providing estimates for the purpose of assessing interrater reliability, however, did not use this structured interview but proceeded with a diagnostic clinical interview however they saw fit.

Procedure

Requirements for participation in the study were briefly explained, informed consent was obtained, and confidentiality was promised. The inventory was then administered to all subjects, either individually to those students randomly sampled, in a group for the large classes, or self-administered to students referred from the Health Service, before subjects participated in a psychiatric interview.

Table 1
Percentages of Interrater Pairs of Psychiatric Ratings as a Function of Degree of Agreement Between Psychiatrists

Degree of agreement	Beck (1970)	Current study
Complete	56	61.5
1 degree of disparity	41	38.5
2 degrees of disparity	2	0
3 degrees of disparity	1	0

Note. The data in column 1 are from *Depression: Causes and treatment* (1st ed.) by Aaron T. Beck, 1970, p. 194. Copyright 1967 by Aaron T. Beck. Reprinted by permission.

Subjects were selected for participation in the psychiatric interview according to depth of depression as measured by the depression inventory, but all interviews were conducted by interviewers blind to depression inventory scores. Structured interviews as described above were conducted to assess the current depth of depression, scored in four levels as none, mild, moderate, or severe according to the criteria developed by Beck for his validation studies (Beck et al., 1961). The interviews were conducted by a board-certified psychiatrist (the third author) with 20 years of experience who maintains both a student practice through the Health Service and an adult practice.

Measures of interrater reliability were obtained at the beginning and end of the study. Both measures were obtained by comparing the primary ratings with those made by one of two additional board-certified psychiatrists, the first (J.H.) with 9 years of clinical experience and the second (A.W.) with 4.

Results

Interrater reliability between the primary psychiatrist providing criterion ratings for depth of depression and the two other psychiatrists, as measured by a Pearson product-moment correlation coefficient, was .62. The proportion of pairs of psychiatric ratings manifesting varying degrees of agreement between psychiatrists, from complete agreement through complete disagreement across the four levels of depression, obtained in the present study and those reported by Beck (1970) are presented in Table 1 for the purposes of comparison.

The Pearson product-moment correlation coefficient between the scores on the depression inventory and the primary psychiatrist's

ratings of depth of depression, with no depression scored as 0, mild as 1, moderate as 2, and severe as 3, was .77 ($SE = .14$, $p < .001$).

For purposes of comparison with Beck's original validation data, the Pearson biserial correlation coefficient was calculated on the same data, after reducing the number of criterion categories from four to two, normal and mild as opposed to moderate and severe. The corresponding coefficient was .79 ($SE = .17$, $p < .001$). Figures summarizing the relationship found in the present study together with those originally reported by Beck (1970) are provided in Table 2.

The dichotomizing of psychiatric ratings for the purpose of calculating a Pearson biserial correlation coefficient produced a rather large disparity between numbers of subjects in the normal and mild category on the one hand and the moderate and severe category on the other. In this case, therefore, this statistic unfortunately yields a large standard error of .17. In the treatment of data originating in this study then, the product-moment correlation, with its smaller standard error, is preferred.

There was no significant association between psychiatric rating of depth of depression and either sex, $\chi^2(3) = .95$, $p = .81$, or grade level, $\chi^2(3) = .85$, $p = .84$.

Beck Depression Inventory means and standard deviations according to the four cate-

Table 2
Correlations Between the Beck Depression Inventory and Psychiatric Rating of Depth of Depression

Study	<i>n</i>	<i>r_b</i>	<i>r</i>	<i>SE</i>	<i>p</i> <
Beck					
Study 1	226	.65		.068	.01*
Study 2	183	.67		.059	.01*
Current	56	.79	.77		
SE		.170	.137		
<i>p</i>		.001	.001		

Note. Data in the first two rows are from *Depression: Causes and treatment* (1st ed.) by Aaron T. Beck, 1970, p. 197. Copyright 1967 by Aaron T. Beck. Reprinted by permission.

* Actual computed probability is less than .001.

gories of psychiatric ratings of depth of depression are furnished in Table 3.

Use of Beck's recommended cutting point of 10 between normal and depressed scores resulted in 2 false positives and 7 false negatives ($p = .18$), whereas retention of his suggested category boundaries produced 27 cases that were inaccurately categorized, of which 15 were overestimates and 12 were underestimates ($p = .70$).

A second, smaller sample ($n = 27$) was run consisting of cases in which a time interval ranging from 1 to 14 days intervened between administration of the depression inventory and psychiatric interview. The Pearson product-moment correlation coefficient between inventory score and psychiatric estimate within this second sample with delay was only .30 ($SE = .21$, $p < .05$). The Pearson biserial correlation coefficient for this same second sample was not meaningful, since the standard error was .36.

Discussion

The primary finding of this study is that the Beck Depression Inventory does indeed appear to be a valid instrument for the measurement of depression in a university population when psychiatric estimate of depth of depression is considered to be the standard. The extent of the correlation is, perhaps, a matter of some surprise, since the pencil-and-paper test was devised by Beck to assess depression conceptualized as a heterogeneous collection of dysfunctions while the psychiatric rating was rendered by a psychiatrist who generally endorses an organic approach to depression. This suggests that depression is a phenomenon of sufficient salience and stability to be detected by various approaches to pathology. When the depression inventory is applied solely as a measure of the current state of depression, the proportions of falsely identified cases clearly indicate that the depression inventory is not associated with any systematic tendency to overestimate or underestimate depression.

There was a strong association between depression inventory scores of 15 and above and a positive response to Item I, which requests

Table 3

Beck Depression Inventory Means and Standard Deviations as a Function of Psychiatric Estimate of Depth of Depression

Statistic	Psychiatric estimate of depth of depression			
	Normal	Mild	Moderate	Severe
<i>n</i>	16	19	17	4
<i>M</i>	3.94	14.10	22.18	19.50
<i>SD</i>	4.46	5.99	8.19	7.55

Note. $n = 56$

information on thoughts of suicide, $\chi^2(1) = 33.65$, $p < .001$, suggesting that a score of 15 or more warrants individual attention to this item. Many subjects, not only those rated as severely depressed by psychiatric estimate, responded positively to the self-harm item.

A major difficulty in this study relates to the question of generalizability, since the criterion measure of psychiatric estimate of depth of depression consists of ratings provided by the primary psychiatrist. Clearly, it would have been desirable to have several psychiatrists participating in the study. The obtained measures of interrater reliability were directed toward this issue, with the results at a generally acceptable level. The somewhat modest correlations between pairs of psychiatric ratings perhaps derive from the fact that the primary and secondary psychiatrists used differing interview procedures, the former following a structured format and the latter using personally determined clinical interviews. The marked contrast between the theoretical orientations of the primary rater and the originator of the depression inventory, however, suggests that a similar correlation between the inventory and the psychiatric estimate might well have been obtained had a number of psychiatrists of varying theoretical persuasions been used to provide criterion ratings.

A second deficiency in this investigation was the small number of subjects who were judged as severely depressed by psychiatric estimate. Sampling procedures were designed to identify this portion of the population, but unfortunately an adequate number was not secured.

The peculiar reversal of the size of the mean scores on the depression inventory between the groups judged moderate and severe by psychiatric estimate shown in Table 3, with the mean for the moderate exceeding that of the severe category, suggests that both the depression inventory and the psychiatric ratings could be invalid at higher levels of depression. The small number of subjects in the severe range, however, indicates that this effect may be attributable only to errors in sampling and precludes drawing further inferences. This remains a question for further study.

The relatively small number of subjects in the moderate and severe categories combined provided little data relevant to the typically reported finding of a significantly higher proportion of females than males undergoing depressions of clinical significance. According to this sample, undetected depressive episodes occurring in the unselected university population are equally likely to afflict both males and females.

The remarkable accuracy of prediction of psychiatric estimate of depth of depression by the depression inventory when psychiatric interview is administered on the same day as the inventory indicates that the inventory elicits little error variance when psychiatric rating is accepted as the criterion. Since in a significant majority of the cases comprising the second sample with time delay (20 out of 27, $p < .025$), the psychiatric rating placed the subject in a milder category of depression than did the depression inventory, it may be inferred from the accuracy of the instrument that in this subsample the depression inventory scores had declined during the intervening time period. This apparent tendency of depression inventory scores to diminish over the passage of time may reflect either spontaneous remission or the subject's attempt to present a more favorable impression during the second assessment.

With a view toward providing a measure of the "steady state," the duration of each symptom of depression that was acknowledged was obtained, with responses scored 1-4 for less than 2 weeks, 2 weeks to less than 6 months, 6 months to any number of years,

and "always," respectively. These duration scores were used as weights by which the item or symptom scores were cross-multiplied, and the total of cross-products provide a composite score representing both severity and duration of depression. It was anticipated that this composite score might not only indicate those subjects who had been depressed but might also predict those who would remain depressed. Though composite scores reflecting self-reported duration correlated very highly with depression inventory scores ($r = .94$, $p < .0001$), they proved less accurate than the "current state" scores as predictors of psychiatric estimate of depth of depression ($r = .40$, $p < .01$). These findings indicate that self-reported duration of symptoms does not provide a useful prognostic index about the future course of depression.

Although the self-reported duration of symptoms failed to provide an accurate measure of steady state depression, the low correlation between scores on the depression inventory and the psychiatric rating for subjects for whom a time interval separated administration of the inventory and the interview raises the issue of the degree to which the state measured by the inventory is stable. These findings suggest the need for longitudinal observations of depressed states in this and other general adult populations to determine their course and outcome.

Reference Note

1. Oliver, J. M., Croghan, L. M., & Katz, N. W. *Prevalence of depression in freshmen and sophomores at four universities as measured in a random sample by the Beck Depression Inventory*. Manuscript submitted for publication, 1976.

References

- American Psychiatric Association. *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, D.C.: APA, in press.
- Beck, A. T. *Depression: Causes and treatment* (1st ed.). Philadelphia: University of Pennsylvania Press, 1970.
- Beck, A. T., & Beck, R. W. Screening depressed patients in family practice: A rapid technique. *Postgraduate Medicine*, 1972, 52, 81-85.
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J.,

- & Erbaugh, J. An inventory for measuring depression. *Archives of General Psychiatry*, 1961, 4, 561-571.
- Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry*, 1972, 26, 57-63.
- Helzer, J. E., et al. Reliability of psychiatric diagnosis: I. A methodological review. *Archives of General Psychiatry*, 1977, 34, 129-133.
- National Institute of Mental Health. *Special report*: 1973. *The depressive disorders*. (Prepared by S. K. Secunda, M. M. Katz, R. J. Friedman, & D. Schuyler. U.S. Department of Health, Education, and Welfare; Health Services and Mental Health Administration, Publ. No. HE 20:2402:D44/4/973.) Washington, D.C.: U.S. Government Printing Office, 1973.
- Seligman, M. E. Fall into helplessness. *Psychology Today*, 1973, 6, 43-48.

Received May 9, 1977 ■

Projective and Repressive Styles of Processing Aversive Information

Alfred B. Heilbrun, Jr.
Emory University

The relationship between projective and repressive styles of dealing with aversive information about the self as well as sex differences in the use of these two defensive styles were studied in normal subjects. An inverse relation between the use of projective attribution and repressive forgetting of negative traits was found independent of sex, supporting the assumption of individual differences in the utilization of the major defense mechanisms. Females used repression more than males, but no sex difference in projection was shown. Implications of these findings for personality theory and research methodology are presented.

Defense mechanisms, the cognitive operations by which people protect themselves from psychological threat, have played a major role in psychodynamic formulations of psychopathology since Freud stressed the centrality of repression (Freud, 1957) and projection (Freud, 1938, 1956) within psychoanalytic theory. A great deal of research has been directed toward verification of repression or projection as clinical constructs or toward the testing of various theories relating to their occurrence or effects. However, reviews of the research literature on repression (Holmes, 1974; Mischel, 1976), the selective forgetting of distressing information, have led to the same pessimistic conclusion. There is little firm evidence for the validity of this construct. The verdict of review concerning the demonstrability of projection as a defense mechanism was somewhat less negative. Holmes (1968) utilized a fourfold typology to consider the research on the attribution of traits to others as a way of denying or otherwise ameliorating the aversive effects of an unacceptable self-characteristic. Was the subject aware of the unacceptable trait or not, and was the attributed trait the same or different from the unacceptable trait? Holmes

concluded that there was no experimental evidence for unconscious projection of unacceptable traits, the classical form of projection proposed by Freud. He did find support for the attribution of traits to others given an awareness of the trait by the individual.

The question remains why it has proven so difficult to develop an impressive backlog of experimental evidence for such popular psychodynamic constructs as repression and projection. The problem may reside in the limitations of theory relevant to the defense mechanisms. It is also possible that the problem is methodological. It is difficult to devise experimental paradigms that are convincing tests of these constructs and that are sufficiently precise so that the results are not vulnerable to alternative explanation.

Yet another possible reason for the equivocal status of repression and projection research is the nature of inference frequently used in such studies. There is a tendency for results to be considered in dichotomous terms; either subjects, considered collectively, demonstrate behavior defined as defensive or they do not. If a sufficient number do not, then the results fail to support the usefulness of the clinical construct. However, defenses as cognitive operations should be expected to occur in varying degrees; for example, some might use repression consistently in dealing with threatening material, others might occa-

Requests for reprints should be sent to Alfred B. Heilbrun, Jr., Department of Psychology, Emory University, Atlanta, Georgia 30322.

sionally engage in repression, and yet others might never deal with threat by using selective forgetting. Furthermore, we might expect individuals to have different hierarchies of defense against threat. One person could be characterized as having a primary repressive style but have other defenses available, whereas another person might maintain primarily a projective style and less frequently deal with threat in other ways.

These observations suggest some methodological innovations in investigating the defense mechanisms. For one, the experimental method should provide the opportunity to assess more than one defensive operation in dealing with the same material. The subject who fails to demonstrate a particular mode of defense (and in previous studies might be thought of as disconfirming that construct) may be shown to be neutralizing threat by a different tactic higher in the hierarchy of defenses. Said another way, this procedure more likely would allow for the demonstration that people have different characteristic styles of dealing with aversive information. Second, the experimental method should allow for the systematic study and analysis of sex differences in characteristic defensive behaviors. The importance of sex differences in ego defense derives first from the fact that they remain virtually unexplored, at least in children (Maccoby & Jacklin, 1974), and, as far as I can discern, in adults as well. If such basic differences exist, it would not only be of some theoretical interest but should be considered in the sampling procedures of subsequent research involving the defense mechanisms.

The present experiment provided subjects with the opportunity to deal with aversive information potentially relevant to themselves by using either a projective or repressive defensive style. No constraints were imposed regarding whether neither, one, or both defenses were used. Two specific issues were investigated. (a) Given the option of either defending against threatening information by projection or repression, will normal subjects tend to use one defensive style and not the other? and (b) Do males and females differ

in the extent to which they use projective and repressive defensive styles?

Method

Subjects

The 47 subjects in this experiment were volunteers from a large undergraduate class at Emory University. Of this number, 30 were males and 17 were females. The mean ages were 18.5 years for the males and 18.3 years for the females.

Measures

The basic instrument from which the evaluative materials were taken for this experiment was the Adjective Check List (Gough & Heilbrun, 1965), a 300-item compilation of behavioral adjectives.

The adjectives on the Adjective Check List have been rated by college students as to their favorability/unfavorability if used to describe a person (Gough, 1955). Seventy-five adjectives have been identified as favorable, 75 as unfavorable, and the remaining 150 adjectives are considered neutral in an evaluative sense. A set of 22 adjectives was selected from each of these three evaluative categories, and these 66 words comprised the critical word list in the present study. Since the projection and repression scores, to be described later, depend on the reliability of these categories, new ratings were obtained from the subjects in this experiment to verify category meanings, since projection and repression scores, to be described later, depend on the reliability of their categories. The mean rating for each evaluative term (given in Table 1) is based on a 7-point scale ranging from highly unfavorable (score = 1) through neutral (neither favorable nor unfavorable—score = 4) to highly favorable (score = 7). Generally speaking, Gough's original groupings held up very well. The favorable and unfavorable categories of evaluative terms, on which both defensive scores were based, provided nonoverlapping values. The rating for each term within these two categories fell on the appropriate qualitative side of the neutral point, and the grand mean rating for the favorable terms (5.84) was over three rating intervals from that for the unfavorable terms (2.46) on a 6-interval scale.

Procedure

Subjects were seen for two sessions. The first was a small group session during which the subjects were asked to rate the favorability or unfavorability of the words comprising the critical word list. These 66 behavioral adjectives were randomly dispersed throughout a rating form made up of 120 adjectives taken from the Adjective Check List. The 54 buffer items were added to reduce a priori familiarity with

Table 1
Favorability Values for Words on the Critical Word List

Qualitative category					
Unfavorable		Neutral		Favorable	
Adjective	M	Adjective	M	Adjective	M
apathetic	2.74	assertive	4.91	adaptable	6.04
careless	2.30	cautious	4.81	cheerful	5.80
cold	1.77	cool	4.13	calm	5.62
cynical	3.19	contented	4.91	capable	6.19
disorderly	2.40	determined	5.68	dependable	6.36
egotistical	2.68	excitable	4.64	enterprising	5.13
fickle	2.60	formal	3.57	frank	5.34
frivolous	3.34	forgetful	2.91	forgiving	6.09
hostile	2.06	hurried	3.32	helpful	6.04
impatient	3.11	informal	5.04	inventive	5.47
indifferent	2.94	idealistic	4.83	imaginative	5.85
intolerant	2.16	initiative	5.51	insightful	5.65
moody	3.49	modest	4.96	mature	6.36
obnoxious	1.72	outspoken	4.06	organized	5.53
prudish	2.49	precise	5.40	patient	5.58
resentful	2.51	reserved	4.23	realistic	5.79
slipshod	2.55	stubborn	3.04	sociable	5.68
selfish	1.91	serious	4.90	sincere	6.34
stingy	2.33	stolid	4.17	stable	5.72
tactless	2.04	trusting	5.89	tolerant	5.45
undependable	1.72	uninhibited	4.83	understanding	6.34
weak	1.94	wary	4.19	warm	6.17

Note. Mean ratings are based on $N = 47$; values range from highly unfavorable (1) to highly favorable (7). The qualitative category is based on Gough's original groupings.

the critical words to be subsequently used in memory tasks.

Subjects were then seen individually in a laboratory session within approximately the next week. During the initial part of the laboratory session, the subject was told that a series of words taken from a survey of mothers who had been asked to describe their college-age children was going to be presented. The subject was instructed to listen carefully and to remember as many of the words on the list as possible. These instructions regarding maternal evaluation, while not made specific to the subject in question because of ethical constraints (i.e., they were not offered as what the subject's mother thought of him or her), have proven successful in creating an evaluative set in several previous studies by me. I (Heilbrun, 1973a) have reviewed these studies elsewhere. The taped list presented the words in repeating triads of favorable, unfavorable, and neutral. The words were spoken in an adult female voice at 5-sec intervals and were presented at a clearly audible volume. Earphones were provided to preclude irrelevant auditory stimulation.

As soon as the final word had been presented, the subject was given a copy of the Adjective Check List and was asked to check those words that had

appeared on the list. The subject was informed that there were 66 words on the list and that exactly that number was to be checked even if it required guessing. Once this was done, the female graduate student experimenter offered the following instructions:

Some of the words used to describe college-age sons (daughters) may be descriptive of you, and some may not. I want you to go through the words you checked and indicate for each one whether you feel it is more descriptive of you than of most college males (females) or whether it is more descriptive of most college males (females) than it is of you.

Subjects then went through the 66 checked words and indicated by using different colored pencils (with a color code continuously available) which of these two alternatives obtained for each term.

Sent to a room down the hall, the subject was met by a second experimenter. This male graduate student administered a filler questionnaire to occupy the subject's time during the next 30 minutes. When the questionnaire had been completed or 25 minutes had elapsed (whichever came first), the subject was told to return to the original room and wait outside in the chair until recalled by the female experi-

menter. This experimenter recalled the subject after exactly 30 minutes had elapsed from the time the subject had left her research room. The purposes of this filler period were twofold: (a) to be sure that sufficient time had elapsed since last exposure to the critical words so that subsequent recognition would qualify as long term and (b) to occupy the subject on a distractor task for as much of this time as possible to avoid rehearsal of the terms previously encountered.

The final part of the laboratory session required the subject to again recognize as many of the words heard on the tape from among those on a new Adjective Check List. Exactly 66 check marks were requested. When this was completed, the subject was thanked and was urged not to discuss the experiment with anyone else.

Results

Projection

Projection was estimated from a score that considered the extent to which the subject attributed unfavorable evaluative terms, correctly recognized from the maternal list, to peers rather than to self. The projection term of this score was *the number of correctly recognized unfavorable terms attributed to others minus the number of correctly recognized unfavorable terms attributed to self*. However, this term, taken alone, would not discriminate between the person who denies unfavorable traits and defensively attributes them to others and the competent individual who justifiably rates peers as more adequately described by the unfavorable terms. Accordingly, a correction term (*the number of correctly recognized favorable terms attributed to self minus the number correctly recognized favorable terms attributed to others*) was subtracted from the projection term to compensate for positive self-esteem. The more positive the difference between the projection and correction terms, the more the selective attribution of negative characteristics to others outweighs the selective endorsement of positive characteristics. In other words, the higher scorer is processing potentially self-relevant negative traits in an extraordinary way relative to positive traits and in a style expected of a projector. Lower scores, whatever they mean about the person, clearly contraindicate a processing of negative traits in a projective style.

This projection scoring procedure was used in a previous study in which theoretically predicted relationships between maternal experience and defensive style were obtained (Heilbrun, 1972). More recently, further evidence of the heuristic value for this index of projection was obtained by showing that valid predictions from a projective test (the Thematic Apperception Test) could be made for projectors but not for nonprojectors (Heilbrun, in press).

A comparison between the mean projection score for males (1.63, $SD = 4.60$) and females (3.00, $SD = 4.90$) revealed no statistically significant difference between the two sexes, $t(45) = .94$, $p > .30$. Males and females did not differ in their utilization of a projective style of processing aversive evaluative information.

Repression

The repression score reflected the extent to which the person failed to correctly recognize unfavorable terms from the critical list during the second retention test after correctly recognizing them during the first retention test and attributing them to himself or herself. In other words, at an earlier time the subject had been aware of aversive self-characteristics but subsequently failed to remember them. Unfavorable terms from the critical list assigned as self-characteristic will be referred to as "repression relevant." Two corrections were required within this score before assuming that the failure to recognize repression-relevant words represented a repressive style of processing negative information.

1. The number of repression-relevant words that the person fails to remember during the second retention test should be considered in terms of the original number of such words. The subject who has self-attributed 10 of the mothers' unfavorable terms and forgets 2 of these (80% recognition) is showing less forgetting than the subject who forgets 2 of these terms from an original number of 4 (50% recognition). Accordingly, a percent retention score (number of repression-relevant terms retained divided by total number of repression-relevant terms) was used to correct

for the size of the original pool of such terms.

2. The number of repression-relevant terms that the person fails to remember should be proportionally greater than would be expected relative to that person's more general long-term recognition rate. This correction was achieved by using all other correctly recognized terms from the initial memory task, those having no theoretical relevance to repression, as a general reference group and determining their retention percentage on the second recognition task. This general reference group included words that were correctly recognized initially and that were unfavorable and attributed to others, neutral and attributed to self or others, or favorable and attributed to self or others.

The final repression score was *percent retention from general reference group minus percent retention from repression-relevant group*. Higher scores in a positive direction suggest repressive forgetting of aversive information, whereas scores near zero or negative scores contraindicate repression and might be taken to suggest sensitization to aversive information.

These percentage correction score distributions for males and females were very similar and approximated normal distributions. The latter observation was confirmed by chi-square goodness-of-fit tests (Maxwell, 1961) applied separately to the distribution of male, $\chi^2(2) = 2.93$, $p > .20$, and female, $\chi^2(2) = 2.23$, $p > .30$, repression scores. Neither distribution deviated statistically from a normal curve.

The projection scores were split independently by sex at their median values (two for

males and three for females), thereby defining high and low projectors. Repression scores (see Table 2) were then analyzed by means of a 2×2 factorial (Level of Projection \times Sex) analysis of variance for unequal cell frequencies (Winer, 1962). There were two significant main effects. Female subjects had higher repression scores than males, $F(1, 43) = 10.48$, $p < .005$, and low projectors had higher repression scores than high projectors, $F(1, 43) = 9.56$, $p < .005$. No Sex \times Projection Level interaction was obtained, $F(1, 43) = 1.02$, $p > .30$.

Control Analyses

Two control analyses of the data were conducted to lend further confidence to the methodology on which the results are based. The first of these concerned the effectiveness of the laboratory analogue that was devised to activate defensive styles in the subjects. Can it be demonstrated that the subjects were responding within an evaluative context in which arousal of defensive behavior might be expected? The instructions indicated that the evaluative terms on the tape came from a survey of mothers asked to describe their college-age sons. The same or very similar instructions have proven effective in eliciting an evaluative set in a host of previous laboratory studies (reviewed in Heilbrun, 1973a) as gauged by the establishment of a theoretically coherent network of results. Effectiveness of the procedures could be shown if the defensive style relationships found in this study when the subjects processed terms ac-

Table 2
Repression Scores as a Function of Level of Projection and Sex

Sex of subject	Level of projection						Com- bined <i>M</i>
	High			Low			
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	
Male	17	-12.13	9.62	13	-4.25	12.04	-8.71
Female	9	-3.70	14.30	8	11.83	12.90	3.61
Combined <i>M</i>		-9.21			1.88		

Note. The higher the scores in the positive direction, the greater the repression. Grand *M* = -4.24.

tually from the tape (and presumably from mothers) could not be replicated for the terms incorrectly "recognized" on the initial memory task and not having a maternal source.

Projection and repression scores were derived for all subjects from the pool of adjectives that had not appeared on the maternal tape but had been checked by each subject as being on the tape. The same statistical analyses were performed as reported earlier in this section. Male ($M = -1.50$) and female ($M = -.47$) projection scores failed to differ, $t(45) = .81$, $p > .40$, as before, and an analysis of variance disclosed no sex difference in the repression scores of males ($M = -15.49$) and females ($M = -23.25$), $F(1, 43) = 1.29$, $p > .25$. Furthermore, no relationship between projection and repression was found, $F(1, 43) = .05$; male high projectors ($M = -14.49$) and low projectors ($M = -16.49$) had virtually identical repression means, and females found to have low and high levels of projection also had very similar means (-26.05 vs. -20.76). Thus, relationships involving defensive styles were obtained only for those terms presented within an evaluative context.

The second control analysis considered the repression score, based on the difference between the retention rate for one group of words comprised totally of unfavorable terms and another group comprised of words that were largely neutral or favorable. It might be contended that the failure to retain the key unfavorable terms for females (or low projectors) is attributable to their more general aversion to all the unfavorable words that appeared in the critical list whether they were considered by the subject to be self-characteristic or not. This failure to recognize negative evaluative terms coming from a maternal source has been demonstrated in a previous experiment using the same words as in the present experiment, similar instructions, and an auditory threshold discrimination procedure (Heilbrun, 1973b). In that study college males showed a specific recognition deficit for negative maternal evaluative terms spoken in a female voice near threshold volume. Examination of the mean number of

adjectives within each qualitative category on the critical list that were correctly recognized on the initial memory task should shed light on this issue. If repression score elevations for females and low projectors are attributable to a more general problem in remembering negative terms having a maternal source, it would be expected that this would also appear in the recognition performance of these two groups immediately following the presentation of the taped list and before the words were sorted into those characterizing the self and others.

Table 3 presents the mean number of recognitions within each qualitative category on the initial memory task for the four Sex \times Projection Level groups. A $2 \times 2 \times 3$ (Level of Projection \times Sex \times Qualitative category) factorial analysis of variance for unequal cell frequencies revealed only one significant effect. Females recognized more critical words ($M = 45.24$) than did males ($M = 40.47$), $F(1, 35) = 5.86$, $p < .025$. However, the absence of a Level of Projection \times Qualitative Category or Sex \times Qualitative Category interaction offered no support for the contention that low projectors or females experienced a more general aversion to unfavorable critical terms.

Discussion

Two basic questions were considered in the present experiment. First, given the option of dealing with unfavorable information about themselves defensively by either projective or repressive styles or nondefensively, will there be a difference in the use of either style between males and females? The results indicated that both sexes used projective attribution of unfavorable characteristics to others to about the same degree. However, females were more prone to use the selective forgetting of unfavorable characteristics involved in repression to a greater extent than males. Since sex differences have received little systematic attention in previous studies of repression, it is difficult to say what the implications of the present results might be for prior research results. They do argue, however, for more careful attention to subject sex in subsequent investigations of repression.

Table 3

Mean Number of Correct Recognitions on the Initial Memory Task for Males and Females Differing in Use of a Projective Defensive Style

Type of subject	Category of evaluative term		
	Favorable	Neutral	Unfavorable
Male			
High projector	13.29	12.47	14.29
Low projector	14.00	12.77	14.23
Female			
High projector	14.11	14.56	15.11
Low projector	16.75	15.00	15.12

The repression results taken alone, even with the sex difference, could have easily taken their place among those of a host of other studies that have sought to demonstrate motivated selective forgetting but, when reviewed as late as 1974, led Holmes (1974) to conclude that "the continued use of repression as an explanation for behavior does not seem justifiable" (p. 651). The overall rate of retention of self-attributed unfavorable terms was only about 4% less than that found for all other combined terms; considering all subjects, this is hardly robust evidence for repression. However, the present investigation went one step beyond its predecessors by offering subjects alternative ways of defending against aversive information. The second question under investigation then became not whether repression (or projection) can be demonstrated but whether, given these alternatives, subjects can be shown to make primary use of one or the other defense. The results suggest that the subjects did tend to use a primary defense style, whether the subject was male or female. Those who projected more usually repressed less, and vice versa. It is worth noting in this context that seven subjects showed no evidence of either a projective or repressive style of defense. This serves as a reminder that many defensive styles are available to cope with threatening information and that the present methodology assessed but two of these. It is also true that the use of either style within this study as a primary defense does not necessarily imply that the same style is the dominant defense in real life. It is the generally inverse rela-

tionship of repression and projection found in this study that is held to be significant to both laboratory investigation and natural behavior.

Holmes (1974) has challenged the interpretation of laboratory studies of repression that have presumably demonstrated the occurrence of this defense mechanism. He considers response competition to be a more parsimonious explanation of the selective loss of threatening material, and two of his own studies (Holmes, 1972; Holmes & Schallow, 1969) encourage this viewpoint. In the earlier of these, the retention of nonevaluative nouns was found to be just as impaired by the introduction of unrelated interpolated material between exposures as by associating word exposure with threats. In the follow-up experiment, a deficit in retention followed not only the association of words with threat but also the association of words with ego-enhancing instructions. It was the cognitive activity associated with either negative or positive evaluation that was thought to be distracting and to interfere with subsequent recall of the words. Holmes' research does demonstrate quite clearly that memory deficits associated with ego threat can be accounted for in response interference terms. Furthermore, his assumption that these deficits are mediated by attentional factors is bolstered by the fact that the impaired recall of terms associated with threat, ego enhancement, or response interference was indistinguishable from control group performance after debriefing and the refocusing of attention on the critical words. Holmes was aware of the inferential problem

of *proving* that the conditions of memory loss for the repression group were the same (or were different) from these for the response interference or ego-enhancing groups, however.

The significance of this inferential problem for the present repression results is two-fold. On the one hand, all subjects were given the same conditions of exposure to the critical words prior to the final recognition task. After initial recognition was measured, each subject judged the relevance of the evaluative terms to self and others and then was maintained on a buffer task for the delay period. Furthermore, repression was assumed only when a specific class of evaluative terms suffered selective loss—those having a maternal source that were unfavorable and assigned as self-characteristic. A response interference interpretation, which assumes the importance of a source of stimulation other than threat, would have to account for why this narrow class of terms suffered more interference than other terms for the same individual. This cannot be reasonably explained in terms of the experimental procedures that were standard for all subjects; it requires an explanation based on the way in which each subject processed the critical terms.

If individual differences in processing negative self-relevant information were invoked as an explanation of selective memory loss, we would be right back to considering repression as a defensive mechanism. It is here that I agree with Holmes in emphasizing response interference and attention. At least as far back as Dollard and Miller's (1950) influential treatise explicating psychoanalytic concepts in learning theory terms, repression has been considered by many as a consequence of anxiety-generated response interference resulting in an attentional shift from threatening material to something more neutral. Given an adequate stimulus, a previous aversive thought is prevented from occurring by the incompatible occurrence of a "stop thinking" response. A similar explanation involving response competition seems to fit the repression results in the present study.

Perhaps one thing that might be accomplished by the present experiment would be to move the investigation of the defense

mechanisms closer to the mainstream of personality research. Rather than concentrating on laboratory demonstration of the defense mechanisms as clinical, trauma-induced phenomena, study of defensive styles could be focused on the ways the normal individual deals with aversive information that threatens self-esteem or emotional well-being. This would broaden our inquiry to a wider range of interesting questions such as the social origins of different defensive styles and the conditions under which defenses are invoked, circumvented, or shifted.

References

- Dollard, J., & Miller, N. E. *Personality and psychotherapy*. New York: McGraw-Hill, 1950.
- Freud, S. Totem and taboo. In A. A. Brill (Ed.), *The basic writings of Sigmund Freud*. New York: Random House, 1938.
- Freud, S. Psychoanalytic notes upon an autobiographical account of a case of paranoia. In, *Collected papers of Sigmund Freud* (Vol. 3). London: Hogarth, 1956.
- Freud, S. On the history of the psycho-analytic movement, 1914. *The complete psychological works of Sigmund Freud* (Vol. 14). London: Hogarth, 1957.
- Gough, H. G. *Reference handbook for the Adjective Check List*. Berkeley, Calif.: Institute of Personality Assessment and Research, University of California, 1955.
- Gough, H. G., & Heilbrun, A. B. *Manual for the Adjective Check List and the Need scales for the ACL*. Palo Alto, Calif.: Consulting Psychologists Press, 1965.
- Heilbrun, A. B. Defensive projection in late adolescents: Implications for a developmental model of paranoid behavior. *Child Development*, 1972, 43, 880-891.
- Heilbrun, A. B. *Aversive material control: A theory of schizophrenic development*. New York: Wiley, 1973. (a)
- Heilbrun, A. B. Recognition of female-spoken and male-spoken evaluative words: Implications for a theory of schizophreniform development. *Journal of Genetic Psychology*, 1973, 123, 297-308. (b)
- Heilbrun, A. B. The influence of defensive styles upon the predictive validity of the Thematic Apperception Test. *Journal of Personality Assessment*, in press.
- Holmes, D. S. Dimensions of projection. *Psychological Bulletin*, 1968, 69, 248-268.
- Holmes, D. Repression or interference? A further investigation. *Journal of Personality and Social Psychology*, 1972, 22, 163-170.

- Holmes, D. S. Investigations of repression: Differential recall of material experimentally or naturally associated with ego threat. *Psychological Bulletin*, 1974, *81*, 632-653.
- Holmes, D. S., & Schallow, J. R. Reduced recall after ego threat: Repression or response competition? *Journal of Personality and Social Psychology*, 1969, *13*, 145-152.
- Maccoby, E. E., & Jacklin, C. N. *The psychology of sex differences*. Stanford, Calif.: Stanford University Press, 1974.
- Maxwell, A. E. *Analyzing qualitative data*. London: Methuen, 1961.
- Mischel, W. *Introduction to personality* (2nd ed.). New York: Holt, Rinehart & Winston, 1976.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.
- Maccoby, E. E., & Jacklin, C. N. *The psychology of*

Received May 25, 1977 ■

Comments

The Right to Treatment: Issues in the Treatment of Homosexuality

Ellie T. Sturgis and Henry E. Adams
University of Georgia

The current controversy regarding the classification and modification of homosexual behavior appears to have been stimulated by pressures on professional organizations exerted by groups such as the Gay Liberation Movement. As a consequence, Davison has argued against the use of sexual reorientation procedures for homosexuality. Davison's recommendation appears motivated by sociopolitical factors rather than psychological evidence, resulting in a neglect of accurate assessment of the needs of the individual. We maintain that the question of the abnormality of homosexuality is an empirical issue, but this issue is irrelevant to the treatment of the individual. The criterion of abnormal behavior is not a necessary prerequisite for behavior modification, since behavior therapists can and do intervene with problems of everyday living to enhance personal effectiveness. The decision concerning the modification of sexual orientation should be based on the life circumstances and values of the individual rather than the value systems of the therapists. We strongly urge clinicians to take a tolerant and impartial view of sexual problems without imposing homosexual, heterosexual, or other value systems on their clients.

Davison (1976), in his discussion of the ethical challenge presented by an examination of homosexual behavior, has raised a number of important and valid issues. However, many of Davison's arguments appear to be based on social and political rhetoric rather than empirical evidence. Such an approach results in a neglect of the assessment of the needs of the individual, which allows therapists to institute, when necessary, an intervention strategy tailored to the requirements of the individual case. If such an assessment is not conducted, the rights of the client may well be violated.

Davison's position that the homosexual be denied the right to modification of sexual orientation violates the right of the individual to treatment and imposes arbitrary values of the clinician in the same manner as the clinician who assumes that all homosexuals should receive treatment for sexual reorientation. The purpose of this discussion is to insure an objective and scientific examination of homosexuality and other target behaviors independent of the current social and political pressures exerted by social activist groups.

What Are Appropriate Target Behaviors?

In his discussion of threats to the ethical foundations of behavior therapy, Davison (1976) challenges the ethics involved in the selection of target behaviors for modification. He cites examples of test anxiety and individuals who lack assertive skills. He questions why therapists treat the "identified client" rather than the systems that produced the anxiety or the individuals who violated the rights of the nonassertive person. A therapist's decision to modify the group behaviors of society rather than behaviors of an individual is legitimate. However, the logical consequence of such a decision implies that the same ethical considerations be applied to society as are applied in the case of an individual. Thus, the intervention must be requested by society, there must exist adequate treatment procedures, and the clinician must know the effectiveness of the treatment program as well as positive and negative side effects.

Particularly relevant to the issue of who should be treated is Davison's (1973) earlier discussion of countercontrol. In that article, he indicated that a therapist cannot force a client (society) to change if that client does not desire to modify a behavior and/or value system. Is society less complex than the individual? The active cooperation

Requests for reprints should be sent to Henry E. Adams, Psychology Clinic, Department of Psychology, University of Georgia, Athens, Georgia 30602.

of the client is required for any treatment method to be effective. Is this not also true for society?

Even though social change is often a desirable goal, it is not particularly evident that behavior therapists have a sufficient data base to initiate or effect such changes. With individual clients, therapy can be objective, empirical, and ethical; however, these values seem to escape us when we discuss society as the client. Unless behavior therapists maintain the same professional values with society that they maintain with individuals, they become political activists rather than psychologists. It is questionable that psychologists interested in social change have any more data for their opinions than do politicians.

Are Therapists Neutral on the Homosexual Issue?

Davison (1976) supports Halleck's (1971) contention that therapists never make ethically neutral decisions. This is a legitimate argument, since research has indicated that the therapists often serve as a modeling and moral agent for the clients (Bandura, 1969). By agreeing to a treatment goal, the clinician conveys the ethical and philosophical approval of the change. The nonneutrality of the therapist is an issue in any treatment procedure. However, Davison implies that the situation is more serious when homosexuality is the issue. Only subjective opinion supports this contention. It is unlikely that therapists are more biased about homosexuality than other problems such as pedophilia, sadism, depression, or schizophrenia. Therefore, to propose that a client not be treated unless the therapist is neutral is to eliminate the helping professions. Value judgments are made about homosexuality as well as for other behavior patterns whether or not society accepts the condition as appropriate. Halleck (1976) contends that although values are always present in a therapeutic situation, the problems are simplified when a comprehensive assessment of the case is conducted. As the circumstances of the behavior pattern vary, it is likely that clinicians will alter their opinions and values accordingly. Halleck argues that a comprehensive assessment increases the likelihood that intervention will not be haphazard. It is further suggested that after the assessment and formulation of any problem behavior regardless of its nature, therapists consult with clients, inform them of their opinions about the problem, discuss possible treatment alternatives, and discuss implications of the alternatives. Included should be a clear statement of the therapist's values that the client can evaluate in accepting or rejecting treatment alternatives. Such an approach decreases the likelihood that a client would unknowingly be influenced by the

value systems of the therapist and also allows the individual to exercise greater personal control over possible consequences of treatment procedures.

Homosexuality: Normality or Abnormality or Is There No Cure Without a Disease?

The issue of abnormality is inevitably raised when psychologists consider the phenomenon of homosexuality. The research data have supported both the "mental health" of the homosexual (Gagnon & Simon, 1973; Green, 1972) and the pathology of the individual (Bieber et al., 1962). Davison, in a telling argument, maintains that before conclusions can be made that etiology factors are pathological, the resultant behavior pattern must first be labeled *abnormal*. Even if a homosexual individual exhibits nonsexual pathological behaviors, it remains to be demonstrated that such behaviors are associated with or are the result of the homosexual condition. Abnormal patterns of behavior observed in homosexuals may be independent of homosexuality and should not be used as a basis for a decision about the abnormality of homosexuality.

The abnormality of homosexual behavior, like the classification of any behavior pattern, is an issue to be resolved empirically rather than through verbal discourse or through the vote of a professional body (Adams, Doster, & Calhoun, 1977). Nevertheless, it does not follow, as Davison suggested, that the abnormality of homosexuality or any other behavior is a necessary or sufficient prerequisite for intervention. Many individuals who seek the aid of a psychologist are normally functioning individuals who experience difficulties in some aspects of their lives. Should the issue of treating individuals who report dissatisfaction with their pattern of sexual preferences be different from treating individuals who are dissatisfied with behavior patterns in nonsexual response systems? When clinicians respond differently to homosexuality than to other problems, are we not reacting to social or political pressures rather than to the basic issue of treatment?

For the clinician, the question of adaptive or maladaptive response patterns should be considered in reference to the individual who requested help. Psychological intervention for maladaptive behavior is warranted when the behavior pattern causes the individual discomfort or distress. However, it does not necessarily follow that behavior therapists should offer assistance only to individuals labeled as abnormal.

Behavioral techniques have often been used with nondeviant behaviors to enhance positive aspects of behavior or increase the personal effectiveness

of an individual. Such target behaviors have included enhancing social skills, assertiveness, study behavior, training in child management procedures, and problem-solving techniques (Goldfried & Davison, 1976). The development of modification techniques does not necessarily indicate that the therapist sees the client as having a "disease" as Davison inferred.

Is Social Prejudice the Primary Agent in Distress?

Davison stated that social prejudice contributes to the problems of the homosexual and causes the individual undue distress. Although this is a valid observation, he implies that social prejudice is unique to homosexuality. On the contrary, social pressures exist for most individuals who deviate from societal norms, regardless of the direction of the deviation. Such individuals are frequently subjected to criticism and discrimination including those who have sought psychological assistance; those of a different race, ethnic, or social group; and even individuals who deviate positively (i.e., high-achieving women or creative individuals; Alper, 1974; Farina & Ring, 1965; Lamy, 1966). The ability of an individual to adapt to such social pressures and to perform adequately is appropriately considered one element of normal functioning. Regardless of the desirability or undesirability of social prejudice, to assume that homosexuals experience more prejudice than others do is questionable. The role of social prejudice in specific deviations (positive or negative) can be more appropriately answered through empirical investigation than through armchair speculation. Until such evidence is available, one cannot conclude that social pressure is the critical factor in the development of distress and desire for change in the homosexual but not in the development of distress and the desire for change in other patterns of behavior.

Davison (1976) stated that society does not attribute problems to heterosexuals on the basis of their sexual behaviors. The validity of this statement is questionable. For example, most rapists are heterosexual, but they are viewed as having problems related to their sexuality. Since the time of Freud, psychological literature has been filled with cases in which the problems or abnormal behavior patterns of heterosexuals have been attributed to fixations at stages of psychosexual development or have resulted from unresolved sexual conflicts. The validity of such hypotheses is irrelevant. However, the impact of Freudian thought on society can neither be estimated nor denied.

Even if society is the causal agent of distress in homosexuality, it does not follow that the solution

is to modify the attitudes of society, as was indicated in our previous discussion of this issue. Although psychologists should try to educate the public concerning the factual issues related to homosexuality (or other behavior patterns), psychologists should refrain from presenting information that lacks a data base. Social propaganda or reeducation efforts that are based on personal opinions rather than empirical evidence can have drastic consequences for society and for the profession. Excellent examples of the danger of such campaigns are the disease concepts of "mental illness" and popularized myths of alcoholism (Rosenhan, 1973; Sobell, Sobell, & Christelman, 1972).

Do Effective Modification Programs Reinforce Societal Beliefs?

Begelman (1975) and Davison (1976) maintain that the very existence of reorientation programs for homosexuality constitutes a significant causal element in reinforcing the doctrine that homosexuality is undesirable. However, the authors fail to substantiate the claim with objective data. A review of recent popular publications concerning behavioral techniques such as Orwell's (1949) 1984 and Burgess' (1962) *A Clockwork Orange* indicates that the public is well aware of the possible dangers of behavioral control, but they are less aware of possible benefits of behavior modification techniques.

A review of the work on "brainwashing" following the Korean War (Brownfield, 1965) indicates that very effective programs were developed that elicited large behavior changes in individuals exposed to these procedures. However, the mere fact that the programs effectively changed attitudes, beliefs, and behaviors of the captives did not cause the public to view the captives as "abnormal." To the contrary, the public outcry was directed at the techniques and the individuals who developed them. More recently, similar public alarm has occurred regarding behavior modification programs in the prisons, schools, and other institutions. Davison has responded to these public criticisms by issuing statements defending behavior therapy to a number of senators, the American Civil Liberties Union, U.S. Department of Health, Education, and Welfare, as well as to a number of magazines and newspapers (Davison, 1974; Davison & Stuart, 1975). If a society considers the target behavior of effective treatment procedures as undesirable, why does society object to the implementation of treatment programs that may be effective with behaviors clearly considered deviant, such as crime? Contrary to Davison's

position, the evidence suggests that society objects to any procedure potentially capable of controlling behavior, regardless of the nature of the behavior. Further, it is doubtful that the availability of techniques encourages their use. The rigorous opposition to behavior modification techniques well illustrates this point.

Are Sexual Preferences Equivalent to Sexual Values?

Davison (1976) seems to assume that sexual preferences of individuals determine their values concerning sexual behavior. The literature on sexual behavior would suggest that this is not always the case. There are numerous examples in which individuals report experiencing annoyance and/or shame concerning patterns of sexual gratification but are unable to bring the patterns under self-control. Such case and group studies involve such sexual patterns as impotence, frigidity, homosexuality, fetishism, transvestism, pedophilia, and sadism (Davison, 1968; Kohlenberg, 1974; LoPiccolo, Stewart, & Watkins, 1972). The problem of incongruent values and preferences is best illustrated by many secondary homosexuals who are married and/or engage in heterosexual encounters but have homosexual arousal. Davison's argument that the homosexual seeks treatment primarily because of social pressures appears to neglect the possibility that there are clients who may actually wish to alter their preference to be congruent with their values rather than changing their value system. The decision to change preferences and/or values should not be made *a priori* but should be determined by aspects of each case.

A Proposal for Therapy with Homosexuals

Davison (1976) and Silverstein (Note 1) have proposed that all individuals with homosexual preferences be treated by desensitizing them to their guilt about homosexual preferences and lifestyles. Such a standardized procedure appears to be a return to a state of affairs in which it was not possible to tailor treatment procedures to the particular needs of the individual's case. Indeed, one distinctive characteristic of behavior modification is that these procedures are appropriate for altering a variety of behaviors. Any *a priori* decision to use only specific techniques eliminates one advantage of behavior therapy that allows the clinician to select treatment procedures appropriate to the individual case. Bieber (1973) has distinguished psychoanalysis and behavior therapy on this very basis.

Critical to the selection of any treatment plan

is a careful assessment of the client's current functioning. The clinician should delineate problem behavior(s) along with possible antecedent (or stimulus) events and the consequent or maintaining factors. A behavioral formulation views homosexual preference as a learned pattern of behaviors that is acquired by individuals through their experiences and that may be maintained or modified in the same way that other behaviors are maintained or modified. Possible targeted behaviors for intervention can include such difficulties as negative attitudes toward the object of sexual preference itself, a desire for sexual reorientation, dissatisfaction with the "gay" lifestyle, interpersonal difficulties in homosexual and/or heterosexual social exchanges, guilt or negative reactions toward the self, sexual dysfunctions with homosexual and/or heterosexual partners, negative attitudes toward members of the opposite sex, or any other personal difficulty that an individual (heterosexual or homosexual) can exhibit. The nature of a problem behavior is not determined when the client states that he or she is homosexual or even that he or she is dissatisfied with the status of being homosexual. Only when a target behavior is thoroughly defined and the antecedents and consequences are determined can the therapist conceptualize the problems and suggest various treatment procedures to the client.

In Davison's (1976) discussion of the termination of preference change programs, he admits that such a move runs the risk of denying treatment to homosexuals who desire an orientation change, not on the basis of social pressures but based on "a sincere desire for things that in our culture are usually part of the heterosexual package—a spouse and children" (Davison, 1976, p. 161). As one solution, he proposes that clinicians accept such risks. This position appears to violate an earlier position adopted by Davison in defense of behavioral treatment for institutionalized individuals. In that article, he stated that "we cannot deny mental patients their right for treatment" (Cashman, 1974, p. 5). How does one justify treatment for institutionalized individuals whose consent, according to Davison, may sometimes be coerced and then refuse to offer treatment for the purpose of sexual reorientation to a homosexual because the decision for treatment may have resulted from social pressure? The decision for treatment should be made by the client rather than the therapist and should be based on an understanding of all relevant aspects of the situation and the client's desires.

In summary, the position taken by Davison appears to be based on philosophical beliefs that

ignore empirical data. His position violates the right of the individual to select treatment goals. The purpose of this article was not to espouse particular views concerning the nature and treatment of homosexuality. To the contrary, the intent was to challenge Davison's position in order to stimulate relevant research and insure protection of the client's right to treatment. Through careful consideration of issues related to the modification of homosexuality, it is hoped that clinicians will arrive at their own conclusions based on rational, objective logic resulting from an evaluation of empirical evidence and the needs of the individual case. As clinical scientists, we can ill afford to make our decisions on the basis of social pressures or personal values. The state of a science cannot be decided by popular vote or verbal mandate. To be effective psychologists, we must ultimately let the data and the requirements of individual cases guide us in the formulation of our views and treatment procedures.

Reference Note

1. Silverstein, C. *Behavior modification and the gay community*. Paper presented at the meeting of the Association for Advancement of Behavior Therapy, New York, October 1972.

References

- Adams, H. E., Doster, J. A., & Calhoun, K. S. A psychological based system of response classification. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Alper, T. G. Achievement motivation in college women: A now-you-see-it-now-you-don't phenomenon. *American Psychologist*, 1974, 29, 194-207.
- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Begelman, D. A. Ethical and legal issues of behavior modification. In M. Hersen, R. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification*. New York: Academic Press, 1975.
- Bieber, I. On behavior therapy: A critique. *Journal of the American Academy of Psychoanalysis*, 1973, 1, 39-52.
- Bieber, I., et al. *Homosexuality: A psychoanalytic study*. New York: Random House, 1962.
- Brownfield, C. A. *Isolation: Clinical and experimental approaches*. New York: Random House, 1965.
- Burgess, J. A. *A clockwork orange*. New York: Norton, 1962.
- Cashman, J. Behaviorists and libertarians: A dialogue on human rights. *Association for Advancement of Behavior Therapy Newsletter*, 1974, 1, 5.
- Davison, G. C. Elimination of a sadistic fantasy by a client-controlled counterconditioning technique. *Journal of Abnormal Psychology*, 1968, 73, 84-90.
- Davison, G. C. Countercontrol in behavior modification. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice*. Champaign, Ill.: Research Press, 1973.
- Davison, G. C. A message from the president. *Association for Advancement of Behavior Therapy Newsletter*, 1974, 1, 1-3.
- Davison, G. C. Homosexuality: The ethical challenge. *Journal of Consulting and Clinical Psychology*, 1976, 44, 157-162.
- Davison, G. C., & Stuart, R. B. Behavior therapy and civil liberties. *American Psychologist*, 1975, 30, 755-763.
- Farina, A., & Ring, K. The influence of perceived mental illness on interpersonal relations. *Journal of Abnormal Psychology*, 1965, 70, 47-51.
- Gagnon, J. H., & Simon, W. *Sexual conduct: The social origins of human sexuality*. Chicago: Aldine, 1973.
- Goldfried, M. R., & Davison, G. C. *Clinical behavior therapy*. New York: Holt, Rinehart & Winston, 1976.
- Green, R. Homosexuality as mental illness. *International Journal of Psychiatry*, 1972, 10, 77-98.
- Halleck, S. L. *The politics of therapy*. New York: Science House, 1971.
- Halleck, S. L. Another response to "Homosexuality: The ethical challenge." *Journal of Consulting and Clinical Psychology*, 1976, 44, 167-170.
- Kohlenberg, R. J. Treatment of a homosexual pedophilic using in vivo desensitization. *Journal of Abnormal Psychology*, 1974, 83, 192-195.
- Lamy, H. F. Social consequences of mental illness. *Journal of Consulting Psychology*, 1966, 30, 450-455.
- LoPiccolo, J., Stewart, R., & Watkins, B. Treatment of erectile failure and ejaculatory incompetence of homosexual ideology. *Journal of Behavior Therapy and Experimental Psychiatry*, 1972, 3, 233-236.
- Orwell, G. 1984. New York: Harcourt, 1949.
- Rosenhan, D. L. On being sane in insane places. *Science*, 1973, 179, 250-258.
- Sobell, L. C., Sobell, M. B., & Christelman, W. C. The myth of "one drink." *Behaviour Research and Therapy*, 1972, 10, 119-123.

Received May 12, 1976 ■

Not Can but Ought: The Treatment of Homosexuality

Gerald C. Davison

State University of New York at Stony Brook

My earlier proposal to terminate change-of-orientation programs rests on moral not empirical grounds. Arguments based on whether therapists can or cannot alter sexual preferences are irrelevant. Therapists, moreover, have no abstract responsibility to accede to requests from clients for certain types of treatment; we work within a host of personal, conceptual, and even legal constraints. Therapists are characterized better as secular priests than as professionals applying ethically neutral techniques. Therapists should attend to large-scale social and political factors in their clients' lives as conscientiously as they attend to intrapsychic and interpersonal variables; our students should study philosophy and politics as well as learning theory and research design. Finally, to urge that therapists desist from sex reorientation programs is not tantamount to exhorting them not to see homosexuals in therapy; indeed, renouncing these widely used programs can help professionals focus on the problems homosexuals (and others) have, rather than on the so-called problem of homosexuality.

I am grateful for the invitation to respond to Sturgis and Adams' (1978) critique of my proposals to terminate change-of-orientation programs for homosexuals (Davison, 1976).

Given the cultural biases against homosexuality, it is problematic to assert that people who ask for change of orientation are expressing a "free wish." We have been remiss in examining why people request certain kinds of treatment (cf. Silverstein, Note 1).

Some of the misunderstanding of my position may stem from a confusion in levels of discourse. Even though I am prepared to argue that *individuals* can benefit from a renunciation of change-of-orientation programs, my proposals are more properly viewed as *institutional* in nature (Rapaport, 1977), that is, as concerned with societal-ethical constraints on human behavior within a framework that is more broad than that familiar to most psychotherapists. Certain issues are better discussed at one level than at another. I believe the issue of therapy for homosexuality should be addressed at an institutional level.

Let me turn now to some of the specific points raised in the critique of Sturgis and Adams

(1978). They are entirely correct to claim that my arguments are based on sociopolitical factors rather than on empirical considerations. They have properly grasped the thrust of my 1976 article, but I do not share their displeasure at my being moved by ethical rather than empirical concerns. What they fail to understand is that the issues we are dealing with as therapists are, indeed, philosophical-ethical ones, and these moral considerations transcend research considerations. To discourse on the empirical level, as they do very well, is simply to misperceive the essence of the issue. Their critique is irrelevant to my article.

Those who continue to offer change-of-orientation treatment to homosexuals do not have a monopoly on sensitivity to clients' rights. I do not believe that the issue can be settled by arguing, as they do, that a therapist has some sort of abstract responsibility to satisfy a client's expressed needs. It is not that simple. As Begelman (1975) has pointed out, therapists constrain themselves in many ways when clients ask for certain kinds of help. There is a host of client requests that therapists do not honor. In fact, the courts (cf. *Kaimowitz v. Michigan Department of Mental Health*, 1973) are becoming involved in denying the "voluntary" requests of patients for certain types of treatment. Requests alone have never been a sufficient criterion for providing therapy.

Clients, moreover, make certain requests of

I wish to thank David A. Begelman for his comments on an earlier draft of this paper.

Requests for reprints should be sent to Gerald C. Davison, Department of Psychology, State University of New York at Stony Brook, Stony Brook, New York 11794.

some therapists and will not do so with others—though, admittedly, it is difficult to collect good data on this. Therapists are purveyors of social ethics, and it is better to own up to this secular priest role (cf. London, 1964) than to continue pretending that it does not exist.

Sturgis and Adams (1978) are correct in asserting that the normality or abnormality of a behavior is irrelevant to whether therapists should try to change it. My earlier discussion of the normality status of homosexuality could have been omitted, but at the time it seemed worthwhile to review both the evidence and the logic behind the futile attempts that have been made to address the issue.

My colleagues do not show an understanding of the social nature of "empirical evidence." Psychologists, like other scientists, do not merely go out and "gather data." They hold preconceived ideas of what they will find and how they will decide they have found it (Davison & Neale, 1978). Scientists adopt paradigmatic ways of defining the problems they will study and how they will study them (Kuhn, 1962). We do not, as Sturgis and Adams (1978) suppose, simply "arrive at . . . conclusions based on rational, objective logic resulting from an evaluation of empirical evidence" (p. 169).

As noted above, their assertion that I based my arguments on social and political considerations is accurate, but to say that *therefore* I am neglecting a careful assessment of the needs of the individual does not follow, certainly not logically or even empirically. The "psychological needs" of a person are seen by Sturgis and Adams as necessarily separate from their social-political needs or pressures. This separation is neither valid nor necessary, and my exhortation is that mental health professionals expand their perspectives to include those sets of variables that are too often overlooked. It is certainly not inherent to a social-learning approach to ignore political and ethical variables; in fact, our general orientation is well suited to this more comprehensive assessment enterprise, as others have creatively demonstrated (e.g., Bandura, 1969; London, 1969; Ullmann & Krasner, 1975).

Sturgis and Adams (1978) contend that "to propose that a client not be treated unless the therapist is neutral is to eliminate the helping professions" (p. 166). This is erroneous on two grounds. First, I am not proposing that clients not be treated unless therapists are neutral. Rather, I am suggesting (after Halleck, 1971) that therapists *cannot* be neutral and that they should realize this. Nowhere, either in what I

have written, taught, or practiced, do I advocate we not do therapy because we are not neutral. Second, to admit to one's biases hardly eliminates the helping professions. If anything, it poses exciting challenges, not the least of which concerns the content of clinical training. Courses in politics, sociology, and philosophy would seem at least as appropriate as courses in learning and statistics.

Having cited Halleck (1971) extensively in my 1976 article, it is not surprising that I generally agree with his comments on that effort (Halleck, 1976). A comprehensive assessment is impossible to argue with, and for therapists to make clear their biases (value systems) is precisely what I am proposing. Along with Sturgis and Adams, I agree with this position. But I would ask only that therapists tell a homosexual that his or her sexual orientation is wrong whenever they embark on a change-of-orientation procedure. The alternative, as stated in my 1976 article, is for therapists to be as vigorous in devising sexual enhancement procedures regardless of orientation as they have been in helping homosexuals become less homosexual.

I have been misunderstood (not necessarily by Sturgis and Adams) to have said or implied that I advocate not treating homosexuals. This is hardly the case. It is one thing to say that one should not treat homosexuality; it is quite another to suggest that one should not treat homosexuals. Indeed, I have urged that therapists *do* finally consider the problems in living that homosexuals really have. Such problems are perhaps especially severe, given the prejudice against their sexual orientation. It would be nice if an alcoholic homosexual, for example, could be helped to reduce his/her drinking without having his/her sexual orientation questioned. It would be nice if a homosexual fearful of interpersonal relationships, or incompetent in them, could be helped without the therapist assuming that homosexuality lies at the root of the problem. It would be nice if a nonorgasmic or impotent homosexual could be helped as a heterosexual would be rather than guiding his/her wishes to change-of-orientation regimens. Implicit in my original argument is the hope that therapists will concentrate their efforts on such *human* problems rather than focusing on the most obvious "maladjustment"—loving members of one's own sex.

Perhaps some people can be hurt by my proposals. To suggest otherwise would be naive. But to assume that people are not being hurt by the

prevalent prejudices is at least as naive.¹ I suppose one has to be reminded of the fact that we live in an imperfect world. I trust there are few therapists who deliberately harm their clients. But it is inherent to my comment that great numbers of people are being hurt by the availability of change-of-orientation programs, and these include individuals who themselves are not seeing therapists. I believe far fewer people would be distressed if we eliminated that option and instead devoted our energies to (a) societal prejudices and (b) to the whole range of problems that homosexuals, heterosexuals, and those in between, have as we all negotiate our way.

I hope the debate continues.

¹ Begelman (1975) takes a different approach in countering the argument that these proposals against sex orientation change are misguided because they may not make people happy. He questions whether the right thing for therapists to do is necessarily what will make their clients happy. Sometimes it is ethical to make a client *more* distressed rather than less so, as when the client's behavior is morally wrong, or when, as is the issue here, the goals of the treatment are ethically questionable.

Reference Note

1. Silverstein, C. *Behavior modification and the gay community*. Paper presented at the annual meeting of the Association for Advancement of Behavior Therapy, New York City, October 1972.

References

- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Begelman, D. A. Ethical and legal issues of behavior modification. In M. Hersen, R. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification*. New York: Academic Press, 1975.
- Davison, G. C. Homosexuality: The ethical challenge. *Journal of Consulting and Clinical Psychology*, 1976, 44, 157-162.
- Davison, G. C., & Neale, J. M. *Abnormal psychology: An experimental clinical approach* (2nd ed.). New York: Wiley, 1978.
- Halleck, S. L. *The politics of therapy*. New York: Science House, 1971.
- Halleck, S. L. Another response to "Homosexuality: The ethical challenge." *Journal of Consulting and Clinical Psychology*, 1976, 44, 167-170.
- Kaimowitz v. Michigan Department of Mental Health, 42 U.S.L. Week 2063 (Mich. Cir. Ct., Wayne Cty., July 10, 1973).
- Kuhn, T. S. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1962.
- London, P. *The modes and morals of psychotherapy*. New York: Holt, Rinehart & Winston, 1964.
- London, P. *Behavior control*. New York: Harper & Row, 1969.
- Rappaport, J. *Community psychology: Values, research and action*. New York: Holt, Rinehart & Winston, 1977.
- Sturgis, E. T., & Adams, H. E. The right to treatment: Issues in the treatment of homosexuality. *Journal of Consulting and Clinical Psychology*, 1978, 46, 165-169.
- Ullmann, L. P., & Krasner, L. *A psychological approach to abnormal behavior* (2nd ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Received April 4, 1976 ■

Brief Reports

Modality, Self-disclosure, and Gender as Determinants of Psychotherapeutic Attraction

Michael R. Kowitt and John P. Garske
Ohio University

The present study investigated the effects of therapy modality and the self-disclosure tendency and gender of the subjects on therapeutic attraction. Forty high and 40 low scorers on a modified self-disclosure questionnaire were asked to rate audiotapes of simulated therapy sessions on several dimensions. The primary results were as follows: High self-disclosers preferred client-centered therapy, whereas low self-disclosers preferred systematic desensitization; client-centered therapy was perceived as providing a greater opportunity for self-exploration, whereas systematic desensitization was perceived as more effective; low-self-disclosing males and high-self-disclosing females rated the therapist as attractive but ineffective; and females were more attracted to systematic desensitization, whereas males were more attracted to client-centered therapy.

Clinical researchers have contended that a subject's attraction to a specific therapeutic modality has a significant impact on the process (Goldstein, 1971) and outcome (Devine & Fernald, 1973) of psychotherapy. The present study investigated this interactive effect in a controlled analogue design. The major hypothesis was that high-self-disclosing subjects would be more attracted to client-centered therapy, whereas low-self-disclosing subjects would be more attracted to systematic desensitization.

College undergraduates (94 males, 84 females) completed Panyard's modified version of the Journal Self-Disclosure Questionnaire and listened to one of two simulated audiotapes of excerpts of psychotherapy sessions. One tape depicted client-centered therapy; the other depicted systematic desensitization. The actors were the same in each (female client, male therapist). Each tape contained segments of the 2nd, 3rd, 8th, and 10th therapy sessions; Sessions 2 and 10 were identical for each to control for the client's report of her problem (interpersonal difficulties) and her evaluation of the effectiveness of the therapy (moderately improved), respectively. The tapes were also matched for duration (10 minutes). Following the tape presentation, each subject rated seven items derived from Goldstein's (1971) Tape Rating

Scale to evaluate the therapist, the technique, and the therapy's effectiveness.

A $2 \times 2 \times 2$ factorial analysis of variance was performed on each of the seven rating variables. The factors were therapy type (client centered/systematic desensitization), subject self-disclosure (upper/lower quartile), and subject sex. Post hoc analyses of the interactions were performed using Cicchetti's procedure. The major finding involved the hypothesized interaction between therapy type and subject self-disclosure ($p < .02$); the low-self-disclosure subjects preferred the therapist who used systematic desensitization ($p < .05$), whereas the high-self-disclosure subjects tended to prefer the client-centered therapist. Low self-disclosers also rated the client-centered therapist as providing less help ($p < .05$). Other significant findings were: Client centered therapy was perceived as providing a greater opportunity for self-exploration, whereas systematic desensitization was perceived as more effective (despite the matched effectiveness of the tapes); low-self-disclosing males and high-self-disclosing females rated both therapists as more attractive but less ineffective; females generally preferred systematic desensitization, whereas males preferred client-centered therapy.

The results revealed that both subject and modality variables independently and interactively affected subject evaluation of psychotherapy. Ostensibly, high self-disclosers were attracted to a modality (client-centered therapy) that matched

Requests for reprints and for an extended report of this study should be sent to John P. Garske, Department of Psychology, Ohio University, Athens, Ohio 45701.

their preference to reveal themselves, whereas low-self-disclosers preferred a modality (systematic desensitization) with structure and direction. There were also complex subject-gender effects. Although these data might be suggestive of differences in actual therapy outcomes (Devine & Fernald, 1973), they should be viewed as preliminary, pending extensions to clinical populations. Moreover, the design was restricted by its use of extreme groups of self-disclosers and a single client-therapist gender pairing.

References

- Devine, D. A., & Fernald, P. S. Outcome effects of receiving a preferred, randomly assigned, or non-preferred therapy. *Journal of Consulting and Clinical Psychology*, 1973, 41, 104-107.
- Goldstein, A. P. *Psychotherapeutic attraction*. New York: Pergamon Press, 1971.

Received February 18, 1976 ■

Factor Analysis of the Ward Atmosphere Scale

Lynn Alden

University of British Columbia, Vancouver, Canada

Eighty-seven inpatient volunteers completed the Ward Atmosphere Scales (WAS) and semantic-differential ratings of their ward. Results of a components analysis did not support Moos's rational grouping of WAS items into 10 dimensions, but they suggest that a global evaluation dimension might underlie 8 of the 10 subscales.

Clinicians are increasingly using measures of the treatment environment in an attempt to identify characteristics of psychiatric treatment programs related to treatment efficiency and effectiveness (Ellsworth & Maroney, 1972; Moos & Schwartz, 1972; Sidman & Moos, 1971). Among the more widely used instruments are a series of measures developed by Moos and his colleagues, including the Ward Atmosphere Scale (WAS; Moos & Houts, 1968) and the Social Climate Survey (Moos, 1968). Because these instruments are gaining widespread acceptance in clinical settings (Milby, Pendergrass & Clarke, 1975; Otto & Moos, 1974), it is important to understand the nature of the information conveyed by the WAS. One question that arises is the extent to which each of the 10 WAS subscales conveys unique information about the environmental unit. Although Moos has not found sufficiently large subscale correlations to warrant collapsing subscales, a recent analysis by Wilkinson (1973) suggested that a single global dimension might underlie environmental measures of this type. Further, the pattern of WAS subscale correlations with the Perception of Ward scales (Ellsworth & Maroney, 1972) and with such factors as patient satisfaction (Houts & Moos, 1969) suggests that a cluster of the WAS subscales may function as a single unit. This study investigated the number and nature of dimensions tapped by the WAS.

This article is based in part on a dissertation submitted in partial fulfillment of the requirements of the PhD at the University of Illinois (Champaign) under the chairmanship of Donald R. Peterson. The author wishes to thank the members of her committee: Donald R. Peterson, Edward Seidman, Stephen Golding, Julian Rappaport, and Ralph Hakstian.

Requests for reprints should be sent to Lynn Alden, Department of Psychology, University of British Columbia, Vancouver, British Columbia Canada V6T 1W5.

Method

Subjects

Subjects were 87 inpatient volunteers at a psychiatric institution for the criminally insane. Subjects ranged in age from 18 to 65, were unmarried (80%), were likely to have a history of previous psychiatric treatment (59%), and were often judged to be violent (71%). The average subject had an eighth-grade education and had been hospitalized for 2-3 years. The most common diagnosis was character disorder (71%), with 8% diagnosed as neurotic and 21% diagnosed as psychotic. Subjects were distributed over eight small wards. Three wards were described as utilizing a conventional milieu approach; three wards followed a behavior modification format; and two wards were described as existential therapy units.

Procedure

All subjects completed the WAS and semantic-differential ratings of their ward. Six adjective pairs were included as part of the semantic-differential technique, with two scales marking each of the three affective components of meaning. The adjective ratings were made on 7-point scales.

Results

The 10 WAS subscales were intercorrelated, subjected to a principal-components analysis with unities in the diagonals and rotated to a varimax criterion of simple structure. Three factors having eigenvalues greater than 1 emerged. All but two subscales (Anger, Staff Control) displayed high positive loadings on the first factor, which accounted for 50% of the total variance. The second factor, accounting for 14% of total variance, was marked by a high positive loading on the Anger scale. The third factor, accounting for 10% of the variance, was marked by Staff Control.

The semantic-differential ratings were subjected to a principal-components analysis following the procedure described above. Two components emerged: evaluation-activity, which accounted for 40% of the total variance, and potency, which accounted for 29% of the total variance.

Factor scores on the WAS and the semantic differential were computed for each subject following the Alberta general factor analysis program (Hakstian & Bay, 1973). Pearson correlation coefficients were computed between the WAS factor scores and the semantic-differential factor scores. Evaluation-activity displayed a moderate relationship, $r(87) = .48$, $p < .01$, with the first component of the WAS, suggesting that the first component reflected subject evaluation of the unit. Low but statistically significant relationships were found between evaluation-activity and the third component, Staff Control, $r(87) = .21$, $p < .05$, and between potency and the second component, Anger, $r(87) = .22$, $p < .05$.

Discussion

The results of the components analysis do not support Moos's rational grouping of WAS items into 10 dimensions. In general, one global dimension was found to underlie subjects' ratings on eight subscales (Involvement, Support, Spontaneity, Autonomy, Practical Orientation, Personal Problem Orientation, Order, and Clarity). The other two components appeared to be specific components, marked by only one scale each. The correlation between the first component of the WAS and the semantic-differential ratings suggests that this component reflects subject evaluation of the unit. It is worth noting that this interpretation supports the conclusions drawn by Wilkinson (1973).

It may be that the WAS functions largely as an attitudinal measure, tapping how positively a

subject feels about the ward. It might prove worthwhile to investigate the extent to which simple client satisfaction measures can achieve the same ends as more extensive measures such as the WAS.

References

- Ellsworth, R., & Maroney, R. Characteristics of psychiatric wards and their effects on patients' adjustment. *Journal of Consulting and Clinical Psychology*, 1972, 39, 436-447.
- Hakstian, A., & Bay, K. *User's manual to accompany the Alberta general factor analysis programme*. Edmonton, Alberta, Canada: Division of Educational Research Services, University of Alberta, 1973.
- Houts, P., & Moos, R. The development of the Ward Initiative Scale for patients. *Journal of Clinical Psychology*, 1969, 25, 319-322.
- Milby, J., Pendergrass, P., & Clarke, C. Token economy versus control ward: A comparison of staff and patient attitudes toward ward environment. *Behavior Therapy*, 1975, 6, 22-29.
- Moos, R. The assessment of the social climates of correctional institutions. *Journal of Research in Crime and Delinquency*, 1968, 5, 174-188.
- Moos, R., & Houts, P. Assessment of the social atmosphere of the psychiatric ward. *Journal of Abnormal Psychology*, 1968, 73, 595-604.
- Moos, R., & Schwartz, J. Treatment environment and treatment outcome. *Journal of Nervous and Mental Disease*, 1972, 154, 264-275.
- Otto, J., & Moos, R. Patient expectations and attendance in community treatment programs. *Community Mental Health Journal*, 1974, 10, 9-15.
- Sidman, J., & Moos, R. On the relation between psychiatric ward atmosphere and helping behavior. *Journal of Clinical Psychology*, 1971, 27, 74-78.
- Wilkinson, L. An assessment of the dimensionality of Moos' Social Climate Scale. *American Journal of Community Psychology*, 1973, 1, 342-350.

Received May 24, 1976 ■

Defensive Externality and Level of Aspiration

Dorothy J. Hochreich
University of Connecticut

This study tested the hypothesis that increasing the salience of achievement cues will result in higher levels of achievement-striving behavior for one subgroup of externals, congruent externals; defensive externals and those with an internal locus of control orientation will behave in a striving manner regardless of level of cue explication. Ninety-six male subjects, defensive externals, congruent externals, and internals, participated in a level of aspiration task under either "game" or "test" instructions. As predicted, congruent externals showed significantly more realistic striving behavior in the test condition ($p < .006$); internals and defensive externals showed similar high levels of realistic striving under both conditions.

Recent research dealing with the relationship between expectancies for internal-external control and achievement behavior (Hochreich, 1975) suggests the utility of using Rotter's interpersonal trust measure to select two groups of externals: defensive (low-trust) externals, who appear to be ambitious, achievement-oriented persons whose endorsement of external statements primarily reflects a verbal defense against failure; and congruent (high-trust) externals, who seem to be less ambitious and less inclined to react to failure by means of blame projection. Previous studies indicate that depending on the particular circumstances, defensive externals will sometimes behave like internals, sometimes like extreme externals.

Lefcourt (1967) suggested that externals, as compared with internals, fail to perceive the availability of achievement rewards unless achievement cues are made quite explicit. Using three levels of cue explication in instructions for a level of aspiration task, he found that externals showed a marked increase in realistic achievement-striving behavior as a result of greater cue explication, but internals did not vary across conditions. If, however, defensive externals are as achievement oriented as internals, they should react in an internal, striving manner whether or not achievement cues are made salient. The less ambitious congruent externals, on the other hand, may require stronger achievement instructions in order to perform successfully.

Three groups of male subjects (32 per group) participated: defensive externals (external and low in trust, determined by median splits); congruent externals (external and high in trust); and internals (with trust scores representing the full range). Within each group, subjects were randomly assigned to participate in Rotter's level of aspiration board task under one of two instructional sets: game condition (task described as a children's game) or test condition (described as a test of motor skills and ability to assess one's own performance). Results were analyzed using Rotter's nine patterns of level of aspiration behavior. Patterns 1-4 represent relatively realistic, responsive, and achievement-striving behaviors. Patterns 5-9 represent less realistic behaviors (overestimation, underestimation, rigidity, or unresponsiveness to feedback).

Overall, subjects showed more realistic patterns in the test condition than in the game condition, $\chi^2(1) = 3.50$, $p < .10$, two-tailed. The three groups did not differ from each other within the test condition (81.5% of defensive externals, 100% of congruent externals, and 87.5% of internals had realistic patterns). A significant difference was found within the game condition, however (88% of defensive externals, 56.5% of congruent externals, and 81% of internals had realistic patterns), $\chi^2(2) = 4.66$, $p < .10$, two-tailed. As predicted, neither internals nor defensive externals varied across conditions (Fisher's exact probability test, $p = .66$ in both cases), whereas congruent externals in the test condition had a significantly higher

Requests for reprints should be sent to Dorothy J. Hochreich, Department of Psychology, University of Connecticut, Storrs, Connecticut 06268.

proportion of realistic patterns than their counterparts in the game condition ($p = .006$, two-tailed). It appears that high levels of cue explication resulted in improved performance for one subgroup of externals, congruent externals. Defensive externals, like internals, attended and reacted to the availability of achievement reinforcements even when cues spelling out such rewards were not present.

References

- Hochreich, D. J. Defensive externality and blame projection following failure. *Journal of Personality and Social Psychology*, 1975, 32, 540-546.
- Lefcourt, H. M. Effects of cue explication upon persons maintaining external control expectancies. *Journal of Personality and Social Psychology*, 1967, 5, 372-378.

Received May 24, 1976 ■

Predictors of Child Noncompliant Behavior in the Home

Rex Forehand, Karen C. Wells, and Ellie T. Sturgis
University of Georgia

The present study examined which data (parent behavior, child behavior, or parent reports) obtained in a clinic setting are the best indicators of child noncompliance in the home. Subjects were 18 mothers and their clinic-referred children. A stepwise multiple regression analysis indicated that two maternal behaviors, beta commands and total rewards, displayed in the clinic were the best predictors of child compliance in the home. The traditionally accepted parent report measures and child behavior in the clinic were not significant predictors of compliance in the home.

Although parent reports and observations of parent-child interactions in a clinic setting traditionally have been used to assess deviant child behavior occurring in the home, many researchers now are using naturalistic home observations. Unfortunately, many clinicians do not perform such assessments due to the time and expense involved. This study examined which assessment procedures in the clinic are the best indicators of child noncompliance, a frequent behavior problem in the home.

Subjects were 18 mother-child pairs. The children, ages 2-9 ($M = 5.1$ years), had been referred for treatment of noncompliance.

Clinic assessment procedures consisted of parent completion of the three scales of the Parent Attitude Test (Cowen, Huser, Beach, & Rappaport, 1970), which pertain to child home behavior problems, and three 20-minute parent-child observations performed over 2 weeks in a playroom. Each observation consisted of 10 minutes in which the mother and child engaged in play activity chosen by the child (free play situation) and 10 minutes in which the mother determined the play activity (command situation). In both situations total (command) rewards, questions, and commands as well as maternal rewards contingent on compliance, as maternal beta commands (vague or interrupted commands to which the child could not comply), and child compliance were recorded. Five 40-minute home observations, performed over 2 weeks, occurred for each mother-child pair. The mother

was instructed to adhere to her daily routine. Child compliance was recorded in the home. Overall reliability, obtained in 20% and 30% of the home and clinic observations, respectively, was 74%.

The 15 independent variables (three questionnaire scales and six behaviors in free play and command situations) and the criterion variable (compliance in the home) were submitted to a stepwise multiple regression analysis. Two maternal behaviors, beta commands in the command situation and total rewards in the free play situation, were significant predictors of child compliance in the home, overall $F(2, 15) = 9.28, p < .01$. The resulting multiple regression equation was $Y = .25166 - .06299 (\text{commands}) + .18623 (\text{rewards})$, indicating a negative relationship between beta commands and compliance and a positive relationship between rewards and compliance. A multiple correlation of .74 indicated that the two independent variables accounted for 55% of the total variance (R^2) in home compliance. A correction for shrinkage, applied to the data to control for the small sample size, reduced the value of R^2 to .51 ($p < .01$). Commands and rewards were responsible for 74% and 26%, respectively, of the predictive ability. The best predictors of child compliance in the home were two clinic-observed maternal behaviors rather than the traditionally accepted parent report measures (questionnaires) or the child behavior (compliance) in the clinic setting.

Reference

- Cowen, E. L., Huser, J., Beach, D. R., & Rappaport, J. Parental perception of young children and their relation to indexes of adjustment. *Journal of Consulting and Clinical Psychology*, 1970, 34, 97-103.

Jorge Mendoza's statistical assistance is gratefully acknowledged.

Requests for reprints and for an extended report of this study should be sent to Rex Forehand, Department of Psychology, University of Georgia, Athens, Georgia 30602.

Received June 14, 1976 ■

MMPI Correlates of Drug Addiction Based on Drug of Choice

Lee Trevithick and Harmon M. Hosch
University of Texas at El Paso

Previous Minnesota Multiphasic Personality Inventory (MMPI) research has focused on addict differences based on substance abused. These investigations have largely failed to detect differences using standard univariate methods. The current study used multivariate analysis to detect differences among groups based on drug of choice (amphetamines, barbiturates, or heroin). Sixty-five addicts (48 males and 17 females) served as subjects. Their composite MMPI profile revealed elements of distress, confusion, and depression as well as sociopathy. Multiple discriminant analysis successfully generated two orthogonal functions that accounted for virtually all of the variance between groups. The loadings of each function were analyzed in terms of the behavioral components characterizing each group. The implications for differential treatment strategies and for theories of personality etiology among drug abusers are discussed.

Previous research (Henriques, Arsenian, Cutter, & Samaraweera, 1972) has failed to detect Minnesota Multiphasic Personality Inventory (MMPI) differences among groups of heroin, barbiturate, and amphetamine abusers. This research, however, relied on a univariate analysis of variance that detects group differences only with respect to each individual scale. Since the MMPI scales are interrelated, the present study used a multiple discriminant analysis (Dixon, 1973) to extract scale combinations that might discriminate among the groups.

Subjects were 65 patients involuntarily committed to a residential drug abuse program. The sample included 48 males and 17 females; mean age was 26.75 years, and the mean years of formal education was 11.6. Although virtually all had a history of multiple drug abuse, each had one preferred drug and would use it first when it was available. Groups thus formed included 29 heroin, 13 barbiturate, and 23 amphetamine abusers.

The composite profile showed elevations ($T > 70$) on the Pd , Sc , D , and Pt scales. This confirms the current view of the addict personality as containing elements of distress, confusion, and depression in addition to sociopathy.

The analysis of MMPI profiles by addict group yielded two discriminant functions that accounted for virtually all of the between-group variance and resulted in 46 cases (70%) being correctly classified into their addict group. It

appears then that the weighted combinations of MMPI scale scores can successfully differentiate the three groups of drug abusers.

Scaled vector weights for the first discriminant function were largely defined by the Hs and Pa scales (characterizing the heroin group) at one pole and by the Hy and Sc scales (barbiturate and amphetamine groups) at the other pole. Tentative labels for the poles of this discriminant function might be egocentric anxiety and acute alienation, respectively.

Thus, although the overall profile distinguishes addicts from nonaddicts, combinations of scales discriminate among the subgroups.

These variations imply that differential treatment strategies might be profitably explored. Intervention might be quite different toward an addict who attempts to distract the therapist from his underlying problems (amphetamine abuser), as compared with the approach toward the nonconformist addict who angrily refuses to cooperate (barbiturate abuser). An intriguing question is whether these profile differences reflect the effects of the drug abused or whether preexisting personality differences resulted in the varying modes of substance abuse. Only longitudinal research can provide the answer to this perplexing dilemma.

References

- Dixon, W. J. *BMD biomedical computer programs*. Berkeley: University of California Press, 1973.
- Henriques, E., Arsenian, J., Cutter, H., & Samaraweera, A. G. Personality characteristics and drug of choice. *International Journal of the Addictions*, 1972, 7, 73-76.

Requests for reprints and for an extended report of this study should be sent to Lee Trevithick, who is now at the Spanish Peaks Mental Health Center, Rural Child Branch, 1083 Lane 25, Pueblo, Colorado 81006.

Received August 13, 1976 ■

Predicting Confabulation from the Graham-Kendall Memory-For-Designs Test

Dan Joslyn

Veterans Administration Hospital
Knoxville, Iowa

John L. Grundvig

Veterans Administration Hospital
Tampa, Florida, and the University of
South Florida College of Medicine

C. J. Chamberlain

Veterans Administration Hospital
Knoxville, Iowa

The Embellishment score on the Memory-For-Designs Test (MFD) may be a useful predictor of confabulation in hospital settings. Embellishments are additional features (lines and curves) that alter the original design. Sixteen long-term hospitalized male psychiatric patients who were judged by staff to be habitual confabulators embellished their drawings on the MFD twice as frequently ($p < .01$) as a group of 16 patients of similar age (40-80) and diagnosis who were judged to be nonconfabulators. This was true for both brain-damaged and chronic schizophrenic patients.

The Memory-For-Designs Test (MFD; Graham & Kendall, 1960) is a frequently used device for the gross screening of brain damage. If performance on a series of standard figures falls below certain normative standards, impairment of cerebral function is inferred. As a consequence of studies involving alternative scoring systems for the MFD (Grundvig, Ajax, & Needham, 1973; Grundvig, Needham, & Ajax, 1970), the clinical impression developed that the embellishment score (Taylor, 1961) was related to confabulatory behavior in patients. The purpose of this study was to determine whether these two characteristics occurred in the same individuals. If such were the case, the Embellishment score on the MFD might be used as a predictor of verbal confabulation in the social world of the patient. The term *confabulation* has been used in a variety of ways ranging from the habit of talking without regard for the truth to outright fabrication, with or without insight. Our definition simply required that

the patient habitually tell improbable stories. No specific attempt was made to rule out intentional deception, exaggeration, or delusion.

Subjects were selected from among chronic-care patient populations in the Nursing Home Care Unit and a psychiatric unit for brain-damaged and chronic schizophrenic patients at the Veterans Administration Hospital, Knoxville, Iowa. Treatment teams on the units were asked to identify patients who habitually told stories of doubtful validity, since patients and staff have had long-standing stable interactions. Sixteen patients about whom there was general staff agreement were selected for the confabulator group (M age = 58.9 years, SD = 9.9), and 16 patients, matched for age, were selected for the nonconfabulator group (M age = 60.7 years, SD = 10.5). Nonconfabulators were not chosen to be representative of nonconfabulators in the general population but of nonconfabulating patients in chronic neuropsychiatric wards. Although not matched for diagnosis, the diagnostic representations in the two groups were comparable in terms of the broad categories of chronic brain syndrome and schizophrenia. Major classifications were as follows: confabulators—10 chronic brain syndrome, 5 schizophrenics, and 1 manic-depressive; nonconfabulators—11 chronic brain syndrome and 5 schizophrenics.

The MFD and Wechsler Memory Scale (Form I) (Wechsler & Stone, 1945) were administered to all patients. The authors who administered and scored the MFD and Wechsler Memory Scale were unaware of patient classification on confabulation, thus ruling out the possibility of subjective

This study was supported by the Medical Research Service of the Veterans Administration.

We thank Emmett Swint, E. J. Luebbbers, Margaret Stevenson, Bernard Schut, Dano Leli, and James Whitehouse for their practical, editorial, and statistical assistance.

Requests for reprints and for an extended report of this study should be sent to Dan Joslyn, Psychology Service, Veterans Administration Hospital, Knoxville, Iowa 50138.

bias. The Wechsler Memory Scale was administered to determine comparability of the two groups in memory functioning. The MFD was scored by the conventional Graham and Kendall (1960) scoring system as well as the modified version (Grundvig et al., 1970) of the Taylor (1961) system. The only score from the modified system that is reported in the present study was the Embellishment score. Each of the 15 designs was inspected for the presence of embellishments, that is, *additional* features that alter the original figure. Only 1 point was scored for each design containing embellishment(s). Therefore, the total possible Embellishment score was 15. To determine the reliability of this score, two of the authors and a graduate student assistant scored the MFD for embellishments. Pairwise reliability coefficients among the three scorers were .85, .54, and .47. The lower two coefficients were obtained from an author whose only exposure to scoring was to study Taylor's (1961) thesis, which gave verbal instructions for scoring embellishments. The first coefficient was obtained between the second author and a graduate assistant whom he had trained personally. Some explicit shaping of criterion classification beyond the reading of Taylor's original definitions and examples seems necessary to improve interscorer reliability. Beyond the computation of reliability coefficients, only the Embellishment scores of the second author were used.

Results show that the confabulators embellished their drawings on the MFD almost twice as frequently ($M = 4.9$, $SD = 3.2$) as the nonconfabulators ($M = 2.5$, $SD = 1.5$), $t(30) = 2.79$, $p < .01$, two-tailed. Furthermore, this relationship was true for both brain-damaged and schizophrenic veterans. The mean Embellishment score for the 10 brain-damaged confabulators ($M = 4.9$) was significantly greater than for the 11 brain-damaged nonconfabulators ($M = 2.4$), $t(19) = 1.99$, $p < .05$, one-tailed. Likewise, the Embellishment score for the 5 schizophrenic confabulators ($M = 5.6$) was significantly greater than that for the 5 schizophrenic nonconfabulators ($M = 2.6$, $t(8) = 2.83$, $p < .02$, one-tailed. In addition to finding these mean differences between groups, we examined the distribution of Embellishment scores to determine the optimal degree of separation between confabulators and nonconfabulators. We obtained a hit rate of 69%.

The groups did not differ significantly in terms of the conventional MFD error score of Graham and Kendall (1960) (for confabulators, $M = 16.7$,

$SD = 14.2$; for nonconfabulators, $M = 13.8$, $SD = 9.0$). These values reflect considerable perceptual-motor impairment. This conventional MFD error score was positively correlated with the Embellishment score, $r(30) = .63$, $p < .001$.

The confabulators and nonconfabulators were practically identical in overall memory functioning as measured by the Wechsler Memory Scale. Means, not age corrected, for the confabulators and nonconfabulators, respectively, were 36.8 ($SD = 11.0$) and 36.2 ($SD = 7.7$). These values corresponded to a Wechsler Memory quotient of 79 and indicated that both groups had significant memory impairment.

In this investigation we began with a somewhat concrete, but hopefully, operational definition of confabulation in the real world. Our impression of a relationship between the tendency to confabulate in the real world and to embellish abstract design reproductions was borne out. Two groups of patients sharing the common characteristics of general memory and perceptual-motor impairment were distinguishable on the basis of a specific type of error production, that is, embellishment. This was true for both brain-damaged and schizophrenic patients and was related to a tendency to confabulate in their real-world interactions with hospital staff personnel. Results suggest that confabulation, as defined in this study, has perceptual, spatial, and motor correlates as well as the well-recognized deficit in verbal behavior.

References

- Graham, F. K., & Kendall, B. S. Memory-For-Designs-Test: Revised scoring manual. *Perceptual and Motor Skills*, 1960, 11, 147-188.
- Grundvig, J. L., Ajax, E. J., & Needham, W. E. Screening organic brain impairment with the Memory-For-Designs Test: Validation of comparison of different scoring systems and exposure times. *Journal of Clinical Psychology*, 1973, 29, 350-354.
- Grundvig, J. L., Needham, W. E., & Ajax, E. J. Comparison of different scoring and administration procedures for the Memory-For-Designs Test. *Journal of Clinical Psychology*, 1970, 26, 353-357.
- Taylor, F. R. A revised scoring system for the Graham-Kendall Memory-For-Designs Test. Unpublished master's thesis, University of Utah, 1961.
- Wechsler, D., & Stone, C. P. A standardized memory scale for clinical use. *Journal of Psychology*, 1945, 19, 87-95.

Received August 17, 1976 ■

Role of Self-instruction and Self-reinforcement in the Modification of Impulsivity

Wilbur J. Nelson, Jr., and John C. Birkimer
University of Louisville

This study sought to determine which components of a previously successful training procedure are necessary in modifying children's impulsivity. Forty-eight impulsive second and third graders of both sexes were used as subjects. Changes in errors and response latencies on Kagan's Matching Familiar Figures Test were used as criteria. Subjects were divided into four groups: (a) self-instruction (SI); (b) self-instruction/self-reinforcement (SI/SR); (c) attention control; and (d) assessment control. Only SI/SR subjects had significantly reduced errors and significantly increased latencies ($p < .001$). Results indicate that self-reinforcement is a necessary aspect of training in the modification of impulsivity.

Training in self-verbalization procedures has been applied to a variety of target behaviors, such as impulsivity (Meichenbaum & Goodman, 1971). In this study, subjects were taught to change their private speech as a strategy for modifying their impulsive behavior. However, little attention was given to the actual components of the self-verbalizations that were used. Thus, both self-instructive (e.g., "draw the line down, down") and self-reinforcing (e.g., "good") components were included. Thoresen and Mahoney (1974) have drawn a distinction between these categories of verbal behavior, thus producing uncertainty as to how the obtained results are to be interpreted.

The purpose of this research was to investigate the effect of self-instruction alone, in addition to the effect of a combination of self-instruction and self-reinforcement, on an impulsive response style. Thus, the design provided a test of whether both components were necessary to produce a change. Since the Matching Familiar Figures Test (MFF) has traditionally been used as a measure in this area, it was used in the present study as a pretest and a posttest. Treatment effects were determined by changes in number of errors committed and latencies to response.

The 48 subjects were black children, both males and females, enrolled in the second and third grades of a public school. An MFF pretest was used to identify the impulsive children from an initial total of 140. Subjects were randomly assigned to each of four training groups. For children in the self-instruction (SI) group, the ver-

balizations were composed of self-instruction alone, whereas subjects in the self-instruction/self-reinforcement (SI/SR) group received a combination of self-instruction and self-reinforcement. The training techniques were similar to those used by Meichenbaum and Goodman (1971). The no-self-verbalization controls received treatment identical to that of the experimental subjects but without self-verbalization training, and the assessment controls were given only the pretest and the posttest.

Simple main effects analyses of variance indicate that SI/SR children made significantly fewer errors, $F(1, 44) = 15.16$, $p < .001$, and had a significant increase in latency to response, $F(1, 44) = 13.20$, $p < .001$, from pretest to posttest MFF. No significant changes were revealed for subjects in any of the other conditions. These results are in agreement with those obtained by Meichenbaum and Goodman (1971). However, the present findings demonstrate that the emission of self-instruction alone does not produce the same results that were obtained in the earlier study. These data provide clear-cut support for the inclusion of self-reinforcement training as a component in cognitive self-instruction packages designed to modify children's impulsivity. Further research is needed to determine whether the self-reinforcement aspect is equally critical in modification of other types of behavior and skills.

References

- Meichenbaum, D., & Goodman, J. Training impulsive children to talk to themselves: A means of developing self-control. *Journal of Abnormal Psychology*, 1971, 77, 115-126.
Thoresen, C. E., & Mahoney, M. J. *Behavioral self-control*. New York: Holt, 1974.

Received September 17, 1976 ■

Requests for reprints and for an extended report of this study should be sent to Wilbur J. Nelson, Jr., who is now at the Mental Hygiene Clinic, Veterans Administration Hospital, Building 5, 555 Wilbard Avenue, Newington, Connecticut 06111.

The Palmar Sweat Index as a Function of Repression-Sensitization and Fear of Dentistry

Charles E. Early and Ronald A. Kleinknecht
Western Washington University

This study investigates the relationship between the repression-sensitization dimension and palmar sweating in response to simulated dental threat stimuli (drill sounds). Sensitizers were found to be more physiologically aroused than were repressors during a brief relaxation period and during a presentation of drill sounds but not during initial or final measurements. Sensitizers also reported being more fearful of dentistry than repressors. These results are seen as consistent with Mischel's social-learning formulation of repression-sensitization in terms of differential attention to threatening stimuli and contrary to repression hypotheses.

Physiological response to threatening situations is highly variable and often shows low to moderate correlations with behavioral or self-report measures. The present study was designed to investigate an individual difference variable, repression-sensitization, which is theoretically related to patterns of response to threat and which might serve to account for variability in physiological response.

Lazarus and Alfert (1964) found consistent with repression hypotheses that repressors, who reported less arousal to threatening movies than did sensitizers, showed greater levels of skin conductance response.

The Palmar Sweat Index (PSI), shown to covary with skin conductance, was used in this study as the measure of physiological arousal in response to simulated dental drill sounds (threat stimulus). Sixty female undergraduates were individually brought into the experimental room where they completed the Byrne Repression-Sensitization (R-S) scale. They were seated in a reclining chair, and the first PSI measure was taken. Next, they listened to a 5-min tape recording of relaxation instructions, designed to reduce arousal associated with the experimental setting, followed by the second PSI. Then on an alternating basis, subjects listened to one of two 15-sec tape recordings: one

with sounds of an actual dentist's drill and one of a child's windup toy car, which simulated the whining sound of a dentist's drill. Subjects were instructed to think of the respective tapes as a dentist's drill or a windup toy. At the end of each tape, the third PSI was taken. Following a brief question period, subjects were told the session was over, and the fourth PSI was taken. Before leaving, subjects completed a 20-item dental fear survey to be correlated with R-S scores.

PSI responses were analyzed using a mixed analysis of variance, with two levels of repression-sensitization (median split) and two sounds (drill or car), with repeated measures at the four points designated above. A significant main effect for repression-sensitization was found showing sensitizers to be more physiologically responsive than repressors, $F(1, 56) = 6.44, p < .05$. Since our interest was in the differential physiological response of repressors and sensitizers to the various conditions, responses were analyzed at each of the four measurement points. This analysis showed sensitizers to be more responsive than repressors during relaxation and during presentations of sounds but not on the initial or final measurement.

A significant trials effect, $F(3, 148) = 17.64, p < .01$, and a Trials \times Stimulus Group interaction, $F(3, 168) = 3.62, p < .05$, were also found. Analysis of the interaction showed that subjects who heard the drill gave greater responses on Trial 3 than on the other trials (which is to be expected, since that was the only trial on which different stimuli were presented).

The fact that sensitizers showed greater physiological responsiveness during both relaxation and threat conditions than repressors is seen as contrary to the repression hypothesis and the Lazarus and Alfert (1964) data. Further, subjects' self-

This research was partially supported by U.S. Public Health Service Grant DE04014, awarded to the second author by the National Institute of Dental Research.

Charles Early is now at the Department of Educational Psychology, Pennsylvania State University.

Requests for reprints should be sent to Ronald Kleinknecht, Department of Psychology, Western Washington University, Bellingham, Washington 98225.

reports appeared to be consistent with their physiological response as evidenced by a moderate but significant correlation of .38 between the dental fear survey and repression-sensitization, suggesting that repressors also report being generally less fearful of dentistry than sensitizers. These data would seem to be in accord with Mischel's (1976) reformulation of the repression-sensitization dimension as reflective of individual differences in learned patterns of attending to or avoiding potentially threatening stimulation, a process that could reasonably be expected to affect physiological responsiveness. An individual difference dimension

such as repression-sensitization might well serve to account for some of the observed variance in physiological responsiveness in fear research.

References

- Lazarus, R. S., & Alfert, E. The short circuiting of threat by experimentally altering cognitive appraisal. *Journal of Abnormal and Social Psychology*, 1964, 69, 195-205.
- Mischel, W. *Introduction to personality*. New York: Holt, Rinehart & Winston, 1976.

Received October 12, 1976 ■

Effects of "Status Sets" on Rotter's Locus of Control Scale

Kay M. Davidson and Kent G. Bailey
Virginia Commonwealth University

Subjects were given sets based on varying levels of social class to determine the susceptibility of Internal-External Locus of Control Scale (I-E) scores to situationally induced frames of reference. College students ($N = 90$) took the I-E scale twice: once while playing the role of a person who had just been hired for one of six ranked occupations and once from their own frame of reference. "In-role" I-E scores revealed the expected positive relationship between levels of status and internality, and these scores differed from the subject's "out-of-role" normal responses. Present results are in accord with recent findings that the I-E scale may be subject to faking and situational effects.

It was recently reported (Deysach, Hiers, & Ross, 1976) that "internal" and "external" instructional sets affected Internal-External Locus of Control Scale (I-E) scores, and, further, scores so obtained differed from those reflecting the subject's own personal beliefs. Add to this the consistent positive empirical relationship between social status and internality (Rotter, Chance, & Phares, 1972), and the question of "faking good" on the I-E scale arises. The present study investigated how well a group of subjects could approximate the empirically established relationship between I-E scores and status. The following hypotheses were evaluated: (a) Groups of subjects given different instructional sets based on varying levels of status can discriminate between items on the I-E scale such that scores will become progressively more internal as status increases; (b) I-E scores obtained under these circumstances will differ from the subject's "normal" scores.

Ninety subjects (65 males and 25 females) were tested, representing six ranked status groups of 15 subjects each. Status levels were derived from a list of 90 occupations previously ranked according to prestige (Hodge, Siegel, & Rossi, 1964). The six jobs, from lowest to highest status, were Level 1, clothes presser; Level 2, grocery checkout clerk; Level 3, bookkeeper; Level 4, assistant personnel manager; Level 5, civil engineer; and Level 6, U.S. representative to Congress. A 50-word description was written about each job to include information about the job holder's salary, educational level, and age. Each subject took the I-E scale while imagining himself or herself in the assigned status role, and immediately thereafter retook the scale from his or her own perspective.

Requests for reprints and for an extended report of this study should be sent to Kay M. Davidson, Department of Psychology, Virginia Commonwealth University, 800 West Franklin Street, Richmond, Virginia 23284.

Analysis of variance revealed that status sets had a strong effect on subjects' I-E scores "while in the role," $F(5, 84) = 26.34$, $p < .001$. I-E scores were clearly inversely related to levels of status. Only Status Level 5 (civil engineer) deviated from a perfect negative relationship, and even then the deviation was not great. Means across the six status levels, going from clothes presser to U.S. representative were 19.50, 17.93, 13.36, 8.53, 9.93, and 5.30. Analysis of variance on the "role" versus normal score differences also proved to be highly significant, $F(5, 84) = 27.43$, $p < .0001$. Difference score means were 11.24, 9.73, 3.86, -4.60, -3.33, and -8.83 across the six status levels, and a multiple comparisons procedure revealed that pairwise combinations were significant at the extremes but not in the middle ranges and adjacent status categories. At extreme status levels, then, a reverberating effect occurred whereby role playing an external role led to internal self-scores and role playing an internal role led to external self-scores.

The above findings indicate that subjects can assume an assigned status role and then produce I-E scores congruent with that role. Further, within-role and normal self-scores tended to differ in a way that suggested a role versus self social-comparison process that raises interesting possibilities for future research.

References

- Deysach, R. E., Hiers, T. G., & Ross, W. A. Situational determinants of performance on the Rotter internal-external locus of control scale. *Journal of Consulting and Clinical Psychology*, 1976, 44, 303.
- Hodge, R. W., Siegel, P. M., & Rossi, P. H. Occupational prestige in the United States, 1925-63. *American Journal of Sociology*, 1964, 70, 286-302.
- Rotter, J. B., Chance, J. E., & Phares, E. J. (Eds.) *Application of a social learning theory of personality*. New York: Holt, Rinehart & Winston, 1972.

Received October 12, 1976 ■

Effects of Model Status and Juvenile Offender Type on the Imitation of Self-reward Criteria

T. John Akamatsu and Parvis A. Farudi
West Virginia University

To explore variables that could facilitate the application of modeling to the treatment of juvenile delinquents, the effects of model status and offender type on the imitation of self-reward criteria were examined. Immature-inadequate and gang-oriented offenders viewed videotapes of models who were either stringent or liberal in self-reward criteria and who were either staff members or peers. Subjects who viewed the liberal model rewarded themselves significantly more than subjects who viewed the stringent model. Significant interactions involving model status and observer type suggest that such factors should be considered in the development of treatment programs.

In the present study the effects of both model status and offender type were examined in an attempt to identify variables that might be of importance in maximizing the effectiveness of modeling approaches with juvenile offenders. The imitation of self-reward criteria was chosen for study because of consistent previous results (e.g., Thelen & Fryrear, 1971) and because this behavior was thought to be relevant to the treatment of juvenile offenders. In addition, Kuncle and Thelen (1972) found that stringent self-reward criteria resulted in improvement on actual task performance, an intriguing finding that warrants replication.

Forty-eight male subjects were exposed to a videotape of either a male adult or peer model who adopted either stringent or liberal self-reward criteria for his performance on a pursuit rotor task. Two types of juvenile offenders were chosen for inclusion in the study based on the system of Quay and Parsons (Note 1): the Behavior Category 1 (BC-1, immature-inadequate) type characterized by inability to cope in a complex world, incompetence, and immaturity; and the Behavior Category 4 (BC-4, socialized-subcultural) type characterized by having bad companions, engaging in gang activities, and being accepted by a delin-

quent subgroup. It was predicted that subjects would imitate the self-reward criteria of the model whom they observed and that model status and offender type would interact. That is, subjects from the immature-inadequate group would be more influenced by the adult model, and subjects from the socialized-subcultural group would be more influenced by the peer model. It was further predicted that exposure to the stringent model would result in better performance on the experimental task.

The procedure was an adaptation of that used by Thelen and Fryrear (1971). Subjects were tested individually, were instructed that they would be participating in an arm-coordination task, and were told that the easiest way to explain how the test worked was to show them a videotaped demonstration. Subjects were then shown one of the four videotapes. After viewing the tape their attention was directed to a chart containing normative information. They were informed that they could take a token that was worth a nickel whenever they thought they deserved it. Subjects received the following predetermined scores on the six test trials: 6, 7, 8, 8, 7, and 9.

The mean number of tokens taken by subjects on trials in which scores of 6 or 7 were obtained was the dependent measure. Preliminary analysis indicated that no significant effects could be attributed to race of subject, so this variable was collapsed. A 2 (stringent-liberal) \times 2 (adult or peer model) \times 2 (behavior category type) analysis of variance was carried out on these data. The analysis revealed a main effect for reward criteria, $F(1, 40) = 41.00$, $p < .0001$. Subjects exposed to the stringent model took fewer rewards

This research was supported by a grant from the West Virginia University Foundation. The authors are grateful to Helene Cavior and Lawrence E. Melamed for their assistance in this project.

Parvis A. Farudi is now at Rusk State Mental Hospital, Rusk, Texas.

Requests for reprints and for an extended report of this study should be sent to T. John Akamatsu, who is now at the Department of Psychology, Kent State University, Kent, Ohio 44242.

($M = .625$) than subjects exposed to the liberal model ($M = 2.33$).

The analysis of variance also revealed a significant Reward Criteria \times Model Status \times Behavior Category Type interaction, $F(1, 40) = 4.122$, $p < .05$. Inspection of the means revealed that as predicted, BC-1 subjects showed greater imitation of the staff models than peer models under both stringent and liberal conditions. The reverse was true for BC-4 subjects. Although the differences were in the predicted directions, post hoc Newman-Keuls analyses revealed that they were not significant. The significant interaction was a function of no difference in the imitation of stringent and liberal peer models by BC-1 subjects. Similar comparisons for BC-4 subjects and for BC-1 subjects who were exposed to staff models were all significant ($p < .01$).

Mean time on target for the six pursuit rotor trials was analyzed with a 2 (stringent vs. liberal) \times 2 (staff vs. peer) \times 2 (BC 1 vs. BC 4) \times 6 (trials) repeated measures analysis of variance, which revealed a main effect for reward criteria. Subjects exposed to the stringent models ($M = 5.70$) performed significantly better than subjects exposed to the liberal models ($M = 4.02$), $F(1, 40) = 6.91$, $p < .02$. A main effect for behavior category type was also detected such that BC-4 subjects performed better than BC-1 subjects ($M = 5.56$ vs. 4.17), $F(1, 40) = 4.73$, $p < .03$. A significant interaction between model status and behavior category type was also detected, $F(1, 40) = 8.122$, $p < .01$. Newman-Keuls analyses revealed that BC-1 subjects who observed the staff model ($M = 3.22$) performed significantly more poorly than BC-1 subjects who observed the peer model ($M = 5.11$, $p < .05$) and BC-4 subjects who observed the staff model ($M = 6.45$, $p < .01$).

The results of the study confirm that self-reward criteria are imitated by juvenile offenders even when less stringent norms are provided. The importance of model status and juvenile offender type in determining the amount of imitation shown is at least partially supported by the data. Differences in imitation as a function of these variables were found as indicated by the significant three-way interaction. Although differences were in the predicted directions, they were not significant. It may be that the restricted range of possible responses (0-3) resulted in a ceiling effect that obscured differences.

A differential susceptibility to staff versus peer models among BC-1 subjects is supported by the data. For these subjects even the potent reward criteria manipulation did not affect imitation of the peer model. It would appear that the staff

model was a more salient source of information for BC-1 subjects than was the peer model.

The results of the analyses on the pursuit rotor scores confirm the finding of Kuncze and Thelen (1972) that observation of a stringent model results in better task performance. Thus, modeling procedures do seem to have effects that go beyond the simple replication of the observed behavior. Greater motivation to perform well, or greater attention to the task at hand, might result from exposure to a stringent model.

The finding of better performance among BC-4 subjects is not too surprising, since on a common-sense level one might expect such subjects to perform better on motor tasks than immature-inadequate types. Involvement by BC-4 types in a broader variety of activities or competitive interaction with peers in sports or games could result in superior performance.

The interaction of model status and behavior category type found in the present study further supports the notion of differential susceptibility to staff and peer model influences among offender types. The BC-1 subjects performed more poorly in the staff model condition perhaps as a function of the perceived discrepancy in abilities between themselves and the model. Such a difference was not found among BC-4 subjects.

In summary, the results of the present study have confirmed the feasibility of the application of modeling techniques to the treatment of juvenile delinquents. Both direct observational learning and indirect motivational effects have been demonstrated. Results concerning the effects of model status and offender type, although not conclusive, suggest that consideration of such variables may be of importance in creating more individualized and maximally effective treatment programs.

Reference Note

1. Quay, H. C., & Parsons, L. B. *The differential behavioral classification of the juvenile offender* (2nd ed.). Robert F. Kennedy Center, Bureau of Prisons, United States Department of Justice, September 1971.

References

- Kuncze, J. T., & Thelen, M. H. Modeled standards of self-reward and observer performance. *Developmental Psychology*, 1972, 7, 153-156.
- Thelen, M. H., & Fryrear, J. L. The effect of observer and model race on the imitation of standards of self-reward. *Developmental Psychology*, 1971, 5, 133-135.

Received October 12, 1976 ■

Factor Structure and Scale Reliabilities of the Adjective Check List Across Time

Stephen H. Miller

Department of Management Sciences
California State University, Hayward

Karlene H. Roberts

School of Business
University of California, Berkeley

Charles A. O'Reilly

Graduate School of Management
University of California, Los Angeles

Carlyle H. Folkins

Department of Psychiatry
University of California, Davis

The purpose of the study was to investigate the factor structure and scale reliabilities of Gough's Adjective Check List (ACL) and to assess their stability over time. Employees in a community mental health center completed the ACL twice, separated by a 1-year interval. After each administration, separate factor analyses were computed. All scales had highly significant test-retest reliabilities. Five factors emerged in each analysis, two of which accounted for about 55% of the common variance. Repetition of factor analysis at two different times resulted in a more stable factor structure than did the usual method of single-time analysis.

The purpose of this research was to investigate the factor structure of Gough's Adjective Check List (ACL; Gough & Heilbrun, 1965) and to assess the stability of this structure over time. Since Gough's original article (Gough, 1960), the ACL has been widely used in personality assessment.¹ Recent studies illustrate the diversity of its application (e.g., Patrick, Zuckerman, & Masterson, 1974; Sowa & Lutter, 1974).

In spite of its popularity, few studies have been done to assess the internal characteristics of the ACL's 24 scales. In factor-analytic investigations, Parker and Megargee (1967) found 4 factors and Scarr (1966) found 10. Neither of the studies provided evidence about the stability of those factors over time. The only available information about scale reliabilities is provided by Gough and Heilbrun (1965).

Respondents were 71 professional employees in a community mental health center. Of the 71, 35 were males. Their average age was 36.2 years, and their average number of years of formal

education was 16.9. Sixty-two of the participants were white, and the rest were distributed across various minority groups. As part of a larger project, the respondents completed the ACL twice, at approximately a 1-year interval.

After each administration, separate factor analyses, using varimax rotations, were computed on the *T* scores for the 24 ACL scales. The two resulting factor structures were compared by calculating a coefficient of congruence (Harman, 1967, pp. 269-272) for each scale score. Test-retest reliabilities were also calculated for each of the 24 scales.

The scales comprising each of the five factors and their loadings on the factors for each administration are given in Table 1. To be included in the factor, the scale had to have a loading greater than an arbitrarily decided on .40 in each of the two analyses. Eighty-one percent of the total communality was accounted for by each of the five factors for each administration. All of the coefficients of congruence were significant beyond the .001 level.

Test-retest correlations (ranging from a low of .51 to a high of .86, $M = .71$) for each of the 24 ACL scales were significant beyond the .001 level. These reliabilities were all greater than those reported in the ACL manual for adult males (6 months between testings) and for medical students (5½ years between testings).

The data clearly indicate the primacy of Factors 1 and 2. Even though Factors 3, 4, and 5 are also replicable and interpretable, they account for less variance in the results than do

Requests for reprints and for an extended report of this study should be sent to Carlyle H. Folkins, University of California, Davis, Sacramento Medical Center, 2315 Stockton Boulevard, Sacramento, California 95817.

¹For a complete Adjective Check List bibliography, see the supplementary bibliographic pamphlet distributed by the Consulting Psychologists Press, 577 College Avenue, Palo Alto, California 94306.

Table 1
Scale Factor Loading for each Administration (t_1 and t_2)

Scale	Factor									
	1		2		3		4		5	
	t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2
Self-confidence	.83	.81								
Achievement	.65	.68			.60	.57				
Dominance	.87	.87								
Exhibition	.74	.72					.52	.45		
Autonomy	.87	.81								
Aggression	.66	.58	-.63	-.71						
Change	.55	.43							.41	.48
Succorance	-.72	-.70								
Abasement	-.95	-.91								
Deference	-.93	-.85								
Defensiveness			.75	.62						
Favorable adjectives checked			.86	.81						
Unfavorable adjectives checked			-.74	-.79						
Self-control			.59	.57						
Personal adjustment			.84	.84						
Intracception			.75	.74						
Nurturance			.85	.80						
Affiliation			.77	.73						
Endurance					.87	.79				
Order					.78	.79				
Heterosexuality							.60	.62		
Counseling readiness							-.64	-.67		
Libility									.47	.63

Note. t_1 = Time 1; t_2 = Time 2.

Factors 1 and 2. Factor 1 seems to reflect "self-confidence and competency." Factor 2 reflects "conventional sociability." Factor 3 appears to reflect an "organized industriousness" coupled with desire for achievement, Factor 4 reflects "lack of anxiety and self-doubt," and Factor 5, "instability or changeability."

The results of the present study do not readily allow comparison with other factor-analytic studies. The samples in previous studies differ, and there is no report in these studies of attempts to assess stability of the factors across time. The Parker and Megargee (1967) study is informative, however, since these authors obtained four factors that are quite similar to the four found here, despite lack of sample similarity and method of administering the instruments. Of the 13 scales in their Factor 1, 4 are present in our Factor 1. Seven of their Factor 1 scales are present in our Factor 2, and 2 are present in our Factor 4. Parker and Megargee's Factor 2 includes 7 scales, all of which are found in our Factor 1. In the Parker and Megargee study, about 66% of the common variance is accounted for by the first two

factors; in the present study, about 55% of the total variance is accounted for by the first two factors.

The data are highly encouraging for users of the ACL. The fact that every scale had a high test-retest reliability increases one's confidence in their validity potential. It appears that whatever the ACL scales measure, they do it consistently. This stability also appears in the factor structure, at least for this sample. If such findings were reported for other samples, the evidence for the utility of the ACL as a reliable measurement device would be enhanced.

The technique of repeated factoring at two different points in time results in a more stable factor structure than does the usual method of single-time analysis. One possible reason for differences in the results reported here and those reported by Parker and Megargee (1967) is that in this study many scales that loaded in the first administration failed to load in the second administration. It would be informative to refactor the Parker and Megargee data using a randomly selected hold-out sample for cross-validation. In light of the many scales that "drop out" during

cross-validation, it would not be surprising to find that such an analysis would provide data more consonant with those reported here. The fact that certain scales did replicate for the same sample across time in this study lends strong support for continuing to cross-validate factor structures for this and other personality diagnostic instruments.

References

- Gough, H. G. The Adjective Check List as a personality assessment research technique. *Psychological Reports*, 1960, 6, 107-122.
- Gough, H. G., & Heilbrun, A. B., Jr. *The Adjective Check List manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1965.
- Harman, H. H. *Modern factor analysis* (2nd ed.). Chicago: University of Chicago Press, 1967.
- Parker, G. V. C., & Megargee, E. I. Factor analytic studies of the Adjective Check List. *Proceedings of the 75th Annual Convention of the American Psychological Association*, 1967, 75, 211-212. (Summary)
- Patrick, A. W., Zuckerman, M., & Masterson, F. An extension of the trait-state distinction from affects to motive measures. *Psychological Reports*, 1974, 34, 1251-1258.
- Scarr, S. The Adjective Check List as a personality assessment technique with children. *Journal of Consulting Psychology*, 1966, 30, 122-128.
- Sowa, P. A., & Lutter, H. S. Attitudes of hospital staff toward alcoholics and drug addicts. *Quarterly Journal of Studies on Alcohol*, 1974, 35, 210-214.

Received November 22, 1976 ■

The Seashore Tonal Memory Test as a Neuropsychological Measure

Carl B. Dodrill and Sureyya S. Dikmen

Department of Neurological Surgery
University of Washington School of Medicine

The sensitivity of the Tonal Memory Test to impaired brain function was evaluated and compared with that of Halstead's Neuropsychological Battery and the Trail Making Test. Neurologic subjects consisted of 102 individuals with histories of head trauma or epilepsy, and control subjects consisted of 68 individuals without histories of neurological problems. In general, the Tonal Memory Test differentiated the normal and neurologic individuals on either a subject-by-subject or group-by-group basis as well as did the other neuropsychological measures, and without excessive overlap with them.

Few studies have been reported that have utilized the Seashore Tonal Memory Test in the investigation of the behavioral correlates of brain lesions. In probably the most extensive report of this type, Milner (1962) used it in evaluating the effects of temporal lobectomy for epilepsy. Although Milner's study and others have suggested that the Tonal Memory Test may be useful in the evaluation of brain-related conditions, none have attempted a formal evaluation of the measure. The goal of this study was to establish the discriminative value of the test in differentiating between normal and neurologic patients and to compare it with the other measures in the neuropsychological battery originated by Halstead and developed by Reitan.

The normal controls included 68 individuals on whom extensive neurological histories revealed no neurological problems or events that may have resulted in such problems. They were solicited from a variety of community agencies and not from any patient population. The mean age of the group was 27.22 years ($SD = 10.07$), and the mean education was 12.28 years ($SD = 1.29$).

This project was supported by National Institutes of Health contract numbers NO1-NS-0-2281 and NO1-NS-2-2326 and Research Grant NS04053 awarded by the National Institute of Neurological and Communicative Disorders and Stroke, U.S. Public Health Service, Department of Health, Education, and Welfare. Sureyya Dikmen is affiliated with the University of Washington Child Development and Mental Retardation Center.

Requests for reprints and for an extended report of this study should be sent to Carl B. Dodrill, Epilepsy Center (ZA-50), Harborview Medical Center, Seattle, Washington 98104.

The neurologic group consisted of 57 individuals with seizure disorders and 45 individuals with head injuries. Each of these subgroups was diversified with respect to type and characteristic of the neurological problem in question, but a positive diagnosis of the problem was made in each instance. The mean age of the total group was 26.45 years ($SD = 10.52$), and the mean education was 12.33 years ($SD = 1.85$).

The Seashore Tonal Memory Test (Seashore, Lewis, & Saetveit, 1960) consists of 30 pairs of tonal sequences including 10 items each of three, four, and five note spans. One note is different in the two sequences of each pair, and the subject must identify that note by number. The final score represents the number of items correctly completed. The other tests in the neuropsychological battery are well-known.

The results of the study are summarized in Table 1. The tests are arranged in this table in order from that producing the largest t value to that producing the smallest. The classification of individual subjects was done on the basis of pre-established cutoff scores for all measures but the Tonal Memory Test. For this measure, a cutoff score between 21 and 22 correct answers produced the largest number of correct individual classifications. Using either the group-by-group or subject-by-subject discriminative procedures, the Tonal Memory Test did as well as did the typical test in the battery in distinguishing between normal and neurologic patients.

Several findings from this study are worthy of comment. First, the Tonal Memory Test appeared to be somewhat better than the Rhythm Test in differentiating between normal and neurologic subjects. This may be due in part to the fact that it is technically more adequate in construction, since

Table 1
Data for Normal and Neurologic Patients on All Test Variables in Order of Efficacy

Test variable	Normal controls ^a		Neurologic patients ^b		<i>t</i>	% correct classification
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Tapping	54.37	5.06	45.64	7.80	8.15**	74
Impairment Index	.16	.18	.47	.30	-7.79**	66
TPT—Localization	5.57	2.22	3.45	2.18	6.17**	69
TPT—Time	11.47	4.67	20.71	12.89	-5.66**	68
Trail Making Test, Part B	59.06	18.35	104.08	73.70	-4.93**	59
Tonal Memory	24.93	4.69	20.29	7.11	4.73**	62
Speech Sounds	4.41	1.98	8.34	6.97	-4.52**	59
Trail Making Test, Part A	24.38	7.16	38.17	24.85	-4.45**	57
TPT—Memory	8.06	1.20	7.00	1.91	4.07**	49
Category	32.84	18.86	47.25	27.06	-3.82*	55
Rhythm	27.00	2.20	25.24	3.44	3.75*	61

Note. TPT = Tactual Performance Test.

^a *n* = 68.

^b *n* = 102.

* *p* < .001 (*t* ≥ 3.35).

** *p* < .0001 (*t* ≥ 3.99).

it has a lower chance score (8) in comparison with the Rhythm Test (15) and therefore a greater range in scores. Its greater difficulty also tends to promote a broader range of scores and greater differentiation between groups.

Although the effectiveness of the Tonal Memory Test in the classification of individual subjects was very modest, we observed that the classification of the other better known tests was usually no better and sometimes somewhat worse (e.g., Category). In addition, in studies in which there was a distinctively better classification of subjects (e.g., Reitan, 1955), we observed that most of the neurologic patients had grossly discernible evidence of tissue damage to the cerebral hemispheres. A detailed evaluation of our patients is not likely to routinely show such grossly discernible alterations in brain structure. They did show very definite signs of impaired brain functions, however. Thus, our neurologic group might be better described as showing signs of "brain impairment" rather than "brain damage," and it is possible that the poorer discrimination by some of the measures found by us may be due, at least in part, to this difference.

With respect to overlap between the Tonal Memory and other tests, the median correlation was .26 for normal controls and .49 for neurologic patients. This appears to be no higher than the correlations of the other tests among themselves.

The apparently superior discriminating ability

of the Tapping Test is explainable at least in part by the fact that 47 patients were taking Dilantin at the time of testing. Dodrill (1975) has shown that this drug has a fairly specific impact on motor performance, and the striking intergroup differences likely thus reflect both impaired brain functions and drug effects.

Overall, the Seashore Tonal Memory Test is a short measure that differentiates normal from neurologically impaired individuals at levels of statistical significance that are consonant with those already used in the neuropsychological battery, and it does so without undue overlap with these measures.

References

- Dodrill, C. B. Diphenylhydantoin serum levels, toxicity, and neuropsychological performance in patients with epilepsy. *Epilepsia*, 1975, 16, 593-600.
- Milner, B. Laterality effects of audition. In V. B. Mountcastle (Ed.), *Interhemispheric relations and cerebral dominance*. Baltimore: Johns Hopkins University Press, 1962.
- Reitan, R. M. An investigation of the validity of Halstead's measures of biological intelligence. *Archives of Neurology and Psychiatry*, 1955, 73, 28-35.
- Seashore, C. B., Lewis, C., & Saetveit, J. G. *Seashore measures of musical talents: Manual*. New York: Psychological Corporation, 1960.

Received November 29, 1976 ■

Relationships Between Dimensions of Anxiety and Sensation Seeking

Barry R. Burkhart, Raymond M. Schwarz, and Samuel B. Green
Auburn University

To determine more precisely the relationships between the general dimensions of sensation seeking and anxiety, 242 undergraduates (130 males, 112 females) completed the Sensation Seeking Scale (SSS) and the S-R Inventory of General Trait Anxiousness (S-R GTA). The intercorrelations among the five scales from the SSS and the four scales from the S-R GTA were computed and compared to theoretical predictions. In general, the empirical findings were consistent with rational and theoretical notions. The majority of the correlations were negative, with the strongest relationship existing between anxiety in physically dangerous situations and sensation-seeking needs. However, the marked variation in the intercorrelations, ranging from moderately negative to low positive, is interpreted as supporting the necessity of multi-dimensional measures of both the anxiety and sensation-seeking constructs.

Recent research has established that a comprehensive theory of motivation must incorporate both tension reduction and stimulus seeking as complementary, multidimensional constructs (McReynolds, 1971). Development of theory and research in this area clearly would be enhanced by the specification of the empirical relationships between the various dimensions of anxiety and stimulus seeking. Segal (1973) reported such data for the Sensation Seeking Scales (SSS; Zuckerman, Note 1) and the S-R Inventory of Anxiousness (Endler, Hunt, & Rosenstein, 1962). As predicted, most of the correlations between the dimensions of sensation seeking and specific anxiety stimulus situations were negative, although they varied greatly in magnitude. The strongest negative relationship occurred between anxiety responses in situations involving physical danger and sensation seeking.

The present study was conducted to extend the results reported by Segal (1973) by attempting to determine the relationships between the general dimensions of sensation seeking and anxiety. The instruments used were the SSS and an extensive revision of the S-R Inventory, the S-R Inventory of General Trait Anxiousness (S-R GTA), which was constructed to rectify the psychometric deficiencies and limited generalizability of the S-R Inventory of Anxiousness (Endler & Okada, 1975). Unlike the older version, which used 11 specific situations, the

S-R GTA is composed of four general stimulus situations: interpersonal, physically dangerous, new or ambiguous, and routine situations. The SSS is composed of five separate factorially derived scales: the General scale, Thrill and Adventure Seeking (TAS), Experience Seeking (ES), Disinhibition (Dis), and Boredom Susceptibility (BS).

The subjects for this study were 242 undergraduates, both males ($n = 130$) and females ($n = 112$), at Auburn University. Each student was asked to complete the SSS and the S-R GTA. Initially, intercorrelations were computed separately for males and females between the four scales of the S-R Inventory and the five SSS scales. In light of the fact that just 1 of the 20 pairs of correlations differed significantly for males and females and due to space limitations, only the correlational matrix for males and females combined is presented.

The results of the correlational analysis are presented in Table 1. The magnitude and the sign of the correlations between the SSS scales and the S-R Inventory subscales varied from moderately negative for Situation 2 (physical danger) to low positive for Situation 4 (innocuous-routine), with Situation 3 (ambiguous-new) and Situation 1 (interpersonal) falling between the two extremes. This ordering of the correlations is rationally appealing. Individuals who prefer to participate in sensation-seeking activities tend not to be anxious in physically dangerous situations. On the other hand, people who are anxious in innocuous, routine situations may be expected to seek sensation-producing experiences. (It should be noted that the positive

Requests for reprints should be sent to Barry R. Burkhart, Department of Psychology, Auburn University, Auburn, Alabama 36830.

Table. 1
Intercorrelations Between Sensation-Seeking Factors and Subscales of the S-R Inventory of General Trait Anxiousness

SSS factor scale	S-R Inventory subscale			
	Interpersonal	Physical danger	Ambiguous	Innocuous
General SSS	-.11	-.30***	-.23***	.18**
Thrill and adventure seeking	-.09	-.37***	-.18**	.10
Experience seeking	-.14*	-.31***	-.23***	.20***
Disinhibition	-.03	-.18**	-.09	.17**
Boredom susceptibility	-.15*	-.22***	-.25***	.24***

Note. SSS = Sensation Seeking Scale.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

relationship between sensation seeking and anxiety arousal in routine situations was significant only for females.) Ambiguous-new situations and interpersonal situations were not as clearly related to sensation seeking, although the former is more similar to the physically dangerous situations than the latter. The correlational pattern supported this continuum.

These results partially replicate the findings of Segal (1973). Again, the majority of the correlations were negative, with the largest ones associated with situations that involved physical danger. However, these results also suggest that differential relationships exist for different dimensions of anxiety and sensation seeking. This theoretically consistent variation makes a compelling case for the necessity of multidimensional measures of both constructs and provides valuable evidence of the construct validity of their measures.

Reference Note

1. Zuckerman, M. *Manual and research report for the Sensation Seeking Scale*. Unpublished manuscript, University of Delaware, 1975.

References

- Endler, N., Hunt, J. McV., & Rosenstein, A. J. An S-R Inventory of Anxiousness. *Psychological Monographs*, 1962, 76(17, Whole No. 536).
- Endler, N. S., & Okada, M. A multidimensional measure of trait anxiety: The S-R Inventory of General Trait Anxiousness. *Journal of Consulting and Clinical Psychology*, 1975, 43, 319-329.
- McReynolds, P. Behavior choice as a function of novelty-seeking and anxiety-avoidance motivations. *Psychological Reports*, 1971, 29, 3-6.
- Segal, B. Sensation seeking and anxiety: Assessment of responses to specific stimulus situations. *Journal of Consulting and Clinical Psychology*, 1973, 41, 135-138.

Received February 22, 1977 ■

Psychotherapy Process Variables Distinguishing the "Inherently Helpful" Person from the Professional Psychotherapist

Beverly Gomes-Schwartz and Joseph M. Schwartz
Vanderbilt University

Therapeutic hours conducted by analytically oriented, experientially oriented, and nonprofessional ("inherently helpful" college professors) therapists were rated along eight process dimensions—therapist exploration, therapist directiveness, feeling attention, task orientation, therapeutic relationship, patient exploration, patient negativism, and patient psychic distress. Between-group differences were obtained on six of the eight dimensions. Only patient negativism and patient exploration failed to yield significant effects. These results substantially replicated the findings of previous analogue investigations.

Although nonprofessional therapists often assume major responsibilities for mental health care, little is known about how the untrained but "inherently helpful" nonprofessional conducts therapy. In simulated interviews, untrained "therapists" tend to make directive interventions rather than exploring patients' feelings or experiences (cf. D'Augelli, Danish, & Brock, 1976). Nonetheless, in similar mock interactions, nonprofessionals (e.g., untrained college students; Pope, Nudler, VonKorff, & McGee, 1974) were as warm, genuine, and empathic as experienced therapists. Indeed, the interviewees felt more accepted and were less anxious in interviews with nonprofessional, as opposed to professional, therapists.

Previous discussions of professional-nonprofessional differences have not taken into account the effects of therapists' theoretical orientations. However, evidence from surveys of therapists' self-described techniques (cf. Sundland & Barker, 1962) and from studies of therapists' responses in analogue interviews (cf. Strupp, 1960) suggests that analytically oriented therapists rely on interpretive techniques, whereas Rogerian or experiential therapists focus on establishing warm, personal relationships.

This research was supported by National Institute of Mental Health Grant 20369. The contributions of H. H. Strupp, principal investigator, and D. Hartley, S. Hadley, and G. Blackwood are gratefully acknowledged.

Requests for reprints and for an extended report of this study should be sent to Beverly Gomes-Schwartz, who is now at Department of Psychology, McLean Hospital, Belmont, Massachusetts 02178.

Although the activities of novice therapists in simulated interactions may be similar to those of nonprofessionals counseling genuinely disturbed clients, generalization of findings from analogue investigations must be viewed with caution. The present study was undertaken to determine whether previous findings would be replicated in the actual therapeutic interactions of analytic, experiential, and untrained therapists.

The patients were 25 unmarried, male college students with elevated scores ($T > 60$) on Minnesota Multiphasic Personality Inventory Scales 2, 7, and 0 who were participating in a psychotherapy outcome study. On a rotational basis, 8 patients had been assigned to analytic therapists (3 male psychiatrists, M experience = 29.0 years), 7 to experiential therapists (2 male psychologists, M experience = 19.5 years), and 10 to alternate therapists. The 6 alternate therapists were experienced male professors (M years since PhD = 18.0) identified by university administrators, faculty, and students as teachers who were frequently approached by students for personal counseling. Although the professors had no formal psychotherapy training, they were recognized in the academic community as inherently helpful people.

Two advanced clinical psychology graduate students independently rated videotapes of each third therapy session on the Vanderbilt Psychotherapy Process Scale (VPPS; Strupp, Hartley & Blackwood, Note 1). This 84-item Likert-type scale was developed from earlier work by Orinsky and Howard (1967) to rate psychotherapy process from the perspective of a clinical observer. Eight subscales derived from the instrument were internally consistent (average co-

efficient $\alpha = .82$) and were rated with satisfactory agreement (average interrater $r = .82$). Two scales tapped therapist factors—Therapist Exploration and Therapist Directiveness. Three scales measured patient dimensions—Patient Exploration, Patient Psychic Distress, and Patient Negativism. Three scales gauged combined patient and therapist contributions—Feeling Attention, Task Orientation and Therapeutic Relationship.

Based on previous analogue findings, the following differences among treatment groups were hypothesized and tested via univariate F tests and Newman-Keuls comparisons ($p < .05$ significance level):

1. Analytic therapists and their patients were expected to engage in greater exploration of psychodynamics than either experiential or alternate dyads. The predicted effect was obtained on ratings of Therapist Exploration, $F(2, 22) = 6.00, p < .01$. Analytic therapists received higher scores than experiential or alternate therapists. Although differences among the groups on Patient Exploration were in the predicted direction, they failed to reach significance, $F(2, 22) = 2.07, p = .15$.

2. Alternate therapists were expected to use more directive interventions (e.g., concrete suggestions) than either professional group. Although overall differences on Therapist Directiveness were significant, $F(2, 22) = 3.56, p < .05$, only analytic therapists were less directive than alternates.

3. Analytic and experiential dyads were expected to focus on important therapeutic issues, particularly examination of patients' feelings, to a greater extent than alternate dyads. On both Task Orientation and Feeling Attention, overall differences were significant, $F(2, 22) = 4.76, p < .05$, and $F(2, 22) = 5.82, p < .01$, respectively. Analytic and experiential pairs received higher scores than alternate pairs on both dimensions.

4. Experiential and alternate therapists were expected to maintain friendlier, more open relationships with their patients than were analytic therapists. Ratings of Therapeutic Relationship yielded a significant overall effect, $F(2, 22) = 10.86, p < .001$, and the predicted differences among groups.

5. No differences among groups in patients' resistance and expression of hostility were anticipated, and no differences on Patient Negativism were obtained ($F < 1$).

6. Patients seen by alternate therapists, as opposed to professionals, were expected to manifest less anxiety in the session. Although a sig-

nificant overall effect was obtained on ratings of Patient Psychic Distress, $F(2, 22) = 4.29, p < .05$, only patients seen by analytic therapists were more distressed during the session than those seen by alternate therapists.

The findings in this study are largely consistent with previous analogue results. Congruent with their theoretical preferences, analytic therapists explored psychodynamics, whereas experiential therapists offered warmth, empathy, and genuineness. In contrast to the professionals who tend to focus on significant therapeutic issues, alternate therapists seemed to engage in informal conversation and advice giving. Perhaps their patients' lesser anxiety in the sessions may be attributed to the alternate therapists' reluctance or inability to examine sensitive intrapsychic or interpersonal issues.

Although the present results suggest that untrained nonprofessionals should not be used with the expectation that they offer the same services as professional therapists, no conclusions as to the effectiveness of the nonprofessionals' interventions can be drawn from these data. Indeed, patients in this study were equally satisfied with their therapist, regardless of the therapist's professional status.

Results of the ongoing Vanderbilt psychotherapy project from which the present data were derived are likely to further clarify the impact of observed variations in therapeutic approach on outcome.

Reference Note

1. Strupp, H. H., Hartley, D., & Blackwood, G. *Vanderbilt Psychotherapy Process Scale*. Unpublished manuscript, Vanderbilt University, 1974.

References

- D'Augelli, A. R., Danish, S. J., & Brock, G. W. Untrained paraprofessionals' verbal helping behavior: Description and implications for training. *American Journal of Community Psychology*, 1976, 4, 275-282.
- Orlinsky, D. E., & Howard, K. I. The good therapy hour. *Archives of General Psychiatry*, 1967, 14, 56-63.
- Pope, B., Nudler, S., VonKorff, M. R., & McGee, J. P. The experienced professional interviewer vs. the complete novice. *Journal of Consulting and Clinical Psychology*, 1974, 42, 680-690.
- Strupp, H. H. *Psychotherapists in action*. New York: Grune & Stratton, 1960.
- Sundland, D., & Barker, E. N. The orientations of psychotherapists. *Journal of Consulting Psychology*, 1962, 26, 201-212.

Received February 24, 1977 ■

"Acceptance," Values, and Therapeutic Change

Larry E. Beutler

Department of Psychiatry
Baylor College of Medicine, Houston, Texas

Stephen Pollack

University of Houston

Avis Jobe

Texas Department of Corrections, Huntsville

This article studied the role of therapist acceptance of patient values, patient acceptance of therapist values, and value persuasion on outcome among 13 psychotherapy dyads. A priori assessment of value acceptance was related to patients' perceptions of their therapists and ratings of improvement, with outcome being enhanced by selective value rejection as well as acceptance. A strong relationship ($p < .01$) was found between the patients' acquisition of their therapists' values and their ratings of improvement.

The current project was designed as an effort to evaluate psychotherapy as a social persuasion process. Hence, self-ratings of improvement and value change were assessed as influenced by (a) the therapist's ability to accept the values of the patient, (b) the patient's ability to accept the therapist's values, and (c) the amount of initial similarity between the patient and the therapist on several value and attitudinal dimensions.

Although the importance of therapist "acceptance" is well recognized in psychotherapy research, little attention has been given either to the role of the patient's ability to accept the therapist or to the possibility that some values are more important to accept than are others. In the current effort to remedy this lack, 13 second-year graduate students in clinical psychology, all having been trained in relationship/insight-oriented therapy, were used as therapists. One case was randomly selected from each therapist's caseload. The patients' ages ranged from 17 to 25 ($M = 20.5$).

Prior to beginning therapy all therapists were asked to complete a series of value questionnaires developed and described elsewhere (Beutler, Jobe, & Elkins, 1974). These scales assessed values relative to others' approval, the threatening nature of the world, God, Communism, Christianity, social laws, and premarital sexual behavior. Each scale was constructed in such a way as to derive latitudes of acceptance and rejection as well as the respondent's preferred attitude on each dimension.

After the first visit and again at the end of 12 psychotherapy sessions, patients were also assessed

on the value questionnaire. On the latter occasion patients completed another questionnaire (Wilson, Morton, & Swanson, Note 1), which was designed to assess their improvement on three dimensions: satisfaction with therapy, satisfaction with the therapist, and global improvement.

Patient and therapist compatibility on the value scale was assessed in two ways. The therapist's values were considered acceptable to the patient if the therapist-preferred position fell within the patient's latitude of acceptance. A similar strategy was used to assess the patient's acceptability to the therapist. In addition, the role of actual similarity as opposed to acceptability was assessed by calculating the number of statements separating the patient's and therapist's preferred positions on each scale.

The influence of three major independent variables on each of three outcome measures was assessed by means of a series of stepwise linear regression analyses. The results suggest that if either the therapist or the patient initially rejected their counterpart's estimate of threat in the world while accepting their views on premarital sex, satisfaction with therapy was increased. The same basic pattern also facilitated global improvement, although only insofar as the therapist's acceptance and rejection of the patient's values were concerned.

A significant correlation ($r = .76$, $p < .01$) was obtained between global improvement ratings and the degree to which patients came to acquire their therapists' attitudes across the seven scales. Apparently, adoption of a therapist's view of life facilitates a patient's sense of positive growth.

Finally, patients' attitudes toward their therapists were enhanced if they rejected their therapists' belief or disbelief in God and acquired their

Requests for reprints should be sent to Larry E. Beutler, Department of Psychiatry, Baylor College of Medicine, Houston, Texas 77030.

values. On the other hand, if the therapists reject their patients' opinions both of Christianity and approval, they also become increasingly attractive to their clients.

Although limited in generalizability, the findings have implication for ethics, training, and the practice of psychotherapy. The concomitant initial rejection of therapists' views of the world with the acceptance of their views about God and sexuality are of particular significance in demonstrating that acceptance need not be complete. Selective rejection of the therapist's values may also be important, although the dynamics of this relationship are unclear. For example, the pervasive relationship between patient-therapist initial disagreement about world threat and outcome is not clear and deserves special research attention.

Apparently, the therapist's attitude toward the patient's values has its greatest impact on the patient's feelings of growth, the patient's attitude

toward the therapist's values seems more strongly related to the development of trust and attraction, and acquiring the therapist's values facilitates general improvement.

Reference Note

1. Wilson, N. C., Morton, W. D., & Swanson, R. M. *Development and validation of a scale to measure treatment effectiveness in psychotherapy*. Paper presented at the meeting of the Rocky Mountain Psychological Association, Denver, May 1971.

Reference

- Beutler, L. E., Jobe, A. M., & Elkins, D. Outcome in group psychotherapy: Using persuasion theory to increase treatment efficiency. *Journal of Consulting and Clinical Psychology*, 1974, 42, 547-553.

Received February 28, 1977 ■

Effect of Physical Attractiveness on Therapists' Initial Judgments of a Person's Self-concept

Stevan E. Hobfoll and Louis A. Penner
University of South Florida

The purpose of this study was to investigate the effect of a person's physical attractiveness on a therapist's initial judgment of that person's self-concept. Videotapes and audiotapes were made of interviews with attractive and unattractive males and females. Graduate students in clinical psychology rated the person they had heard (seen) as to self-concept. As hypothesized, physically attractive persons of both sexes were rated as having better self-concepts than unattractive persons. Further, the self-concept ratings of attractive females increased significantly from the audiotape to the videotape conditions, whereas the ratings of all the other stimulus persons remained the same.

Social psychological research on physical attractiveness has shown that a physically attractive stimulus person will be seen as possessing more positive personality attributes than a physically unattractive person (cf. Berscheid & Walster, 1974). Other studies have suggested that this stereotype may be more pronounced for female than for male stimulus persons. Despite the growing concern with sex role stereotypes as they may affect clinical practices (Report of the Task Force, 1975), there has been relatively little research conducted on the extent to which clinicians adhere to these societal stereotypes in their judgments of their clients.

The purpose of this study was to examine the relationship between a person's (judged) physical attractiveness and a therapist's estimate of that person's self-concept. It was hypothesized that ratings of a stimulus person's self-concept would be positively related to the stimulus person's physical attractiveness and that this relationship would be stronger for female than for male stimulus persons.

Undergraduates rated photographs of 83 of their classmates on an 11-point scale of physical attractiveness. On the basis of these ratings, the 2 most attractive males, 2 least attractive males, and a similar number of females were selected as the stimulus persons. These 8 people were interviewed, and a 10-minute audiotape or videotape of the interview was presented to 13 male

and 3 female graduate students in clinical psychology, each of whom had had at least 1 year of clinical experience. After hearing (viewing) an interview with a stimulus person, the student clinicians rated that person on a 7-point scale of self-concept. Data were analyzed via two 8×8 Latin squares, with mode of presentation as the between-subjects variable and physical attractiveness, sex, and tapes (2 stimulus persons were used in each condition) as the within-subjects variables.

Stimulus persons rated as most attractive were judged by student clinicians as having a significantly better self-concept than unattractive stimulus persons, $F(1, 91) = 349.59, p < .001$, and males were rated as having better self-concepts than females, $F(1, 91) = 8.29, p < .01$. There was a significant Mode of Presentation \times Attractiveness interaction, $F(1, 91) = 8.24, p < .01$. Although the attractive stimulus persons were rated as having better self-concepts than the unattractive people in both the audiotape and videotape conditions, the self-concept ratings of attractive persons increased significantly from the audiotape to the videotape conditions. The ratings for unattractive stimulus persons remained the same. Finally, there was a significant Mode of Presentation \times Sex \times Attractiveness interaction, $F(1, 91) = 6.309, p < .05$. Attractive females were rated as having significantly better self-concepts in the videotape condition than in the audiotape condition, but ratings of all other stimulus persons did not change.

The finding that attractive stimulus persons received higher self-concept ratings in the audiotape condition was unexpected. Explanations of this finding may lie in Berscheid and Walster's

Requests for reprints and for an extended report of this study should be sent to Louis A. Penner, Department of Psychology, University of South Florida, Tampa, Florida 33620.

(1974) argument that people judged by others as physically attractive may have more positive social experiences than unattractive persons. Thus, the higher self-concept ratings of attractive stimulus persons in the audiotape condition may have been due to these people manifesting more self-confidence and better social skills in the interview than unattractive stimulus persons. At the same time, the results of this study suggest that the student clinicians did adhere to societal stereotypes as to the relationship between physical attractiveness and the possession of positive attributes. Further, it would appear that these stereotypes exerted their

strongest influence when the stimulus person was an attractive female.

References

- Berscheid, E., & Walster, E. Physical attractiveness. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 7). New York: Academic Press, 1974.
- Report of the task force on sex bias and sex-role stereotyping in psychotherapeutic practice. *American Psychologist*, 1975, 30, 1169-1175.

Received March 4, 1977 ■

Empathy and Imagery in Avoidance Behavior Reduction

John T. Esse
University of Miami

Wallace Wilkins
University of Maine at Orono

This therapy analogue study was designed to assess the relative effects of therapist empathy and instructed imagination of hierarchy scenes on avoidance behavior reduction. Although the communication of differential therapist empathy was validated, behavior change attributable to therapist empathy was minor in comparison to the effects of imagery instructions. Imagery instructions delivered in a relatively unempathetic fashion produced as much avoidance reduction as imagery instructions delivered in an empathetic manner. Unempathetic imagery instructions also produced significantly greater avoidance reduction than the establishment of an empathetic relationship without instructed imagery exercises.

Although the efficacy of systematic desensitization, as a treatment package, has been documented, considerable controversy exists about the specific elements and the theoretical mechanisms that account for gain. Therapist empathy, defined as an ability to understand a client's experiences and communicate that understanding to the client, is also a focus of considerable debate concerning the conditions that facilitate therapeutic gain. In this study, to assess the relative contributions of imagery instructions and therapist empathy, imagery exercises were delivered with and without empathy; empathy was delivered with and without imagery instructions.

The subjects were 30 undergraduate female students who indicated "much fear," "very much fear," or "terror" on Item 39 (snakes) during a classroom-administered Fear Survey Schedule II (Geer, 1965), and who were unable to touch with bare hands a live, 3-foot (.91 m) king snake during the Behavioral Avoidance Test (BAT; Nawas, 1971). All participants in the study received credit toward partial fulfillment of introductory psychology course requirements.

The pre-BAT and post-BAT were conducted by a female experimenter, who remained experimentally blind as to the hypotheses and design of the study. Following the pre-BAT, subjects were contacted for a treatment session that lasted for a maximum of 45 minutes. All sessions were administered individually by a female graduate student after subjects were informed that "the procedure

which you are about to undergo has often been found helpful in reducing snake fear." Immediately following the treatment session, each subject completed a set of ratings, which included the 16 empathetic understanding items from the Barrett-Lennard Relationship Inventory (Barrett-Lennard, 1962), and was tested on the post-BAT.

Mechanical Imagery Procedure

The 10 members of this group received a modification of the instructed attention shift procedure, shown by Wilkins and Domitor (1973) to be effective in reducing avoidance behavior in a short period of time. The procedure was executed in a mechanical fashion that allowed little deviation from the predetermined procedure and minimized therapist empathetic reactions. The hierarchy of scenes that was presented for instructed imagination consisted of the 20 BAT steps read in the second-person present tense beginning with the least threatening and proceeding to the most threatening scene.

Empathetic Imagery Procedure

The 10 members of this group received the basic procedures of the mechanical imagery procedure (MIP) delivered in a more informal, responsive, and spontaneous manner. During the presentation of the imaginal scenes, the therapist varied the wording of the scene descriptions and the tone of her voice so as to be more responsive to the feelings that the subject appeared to be experiencing.

Empathetic Conversational Procedure

The 10 members of this group were asked to talk about feelings about snakes. There was verbal

Requests for reprints and for an extended report of this study should be sent to John Esse, who is now at Trend Community Mental Health Services, 242-B Second Avenue East, Hendersonville, North Carolina 28739.

exploration of any methods that the subject felt would be helpful in reducing her fear, and, at some point during the session, she was also asked to imagine herself approaching a snake and to describe the feelings that were aroused. Beyond this general suggestion, however, no directives were given as to what she could do to become less fearful. During the entire session the therapist attempted to be maximally sensitive to and reflective of emotional expressions.

Validation of Empathy Manipulation

Within a possible range from -48 to $+48$, mean ratings of empathetic understanding for the MIP, the empathetic imagery procedure (EIP), and empathetic conversational procedure (ECP) groups were, respectively, 1.70 ($SD = 6.147$), 21.90 ($SD = 5.466$), and 18.30 ($SD = 7.602$). An examination of specific comparisons among group means showed that significantly greater therapist empathy was communicated during the EIP than during the MIP $F(1, 18) = 60.301$, $p < .001$, and that significantly greater empathy was communicated during the ECP than during the MIP $F(1, 18) = 28.831$, $p < .001$. Statistically nonsignificant differences occurred between the EIP and the ECP groups.

Avoidance Behavior Change

Mean pretreatment BAT scores for the MIP, EIP, and ECP groups were, respectively, 8.20 ($SD = 3.259$), 8.10 ($SD = 2.378$), and 8.00 ($SD = 3.266$). Posttreatment BAT means were, respectively, 13.90 ($SD = 3.573$), 13.30 ($SD = 2.263$), and 10.10 ($SD = 5.259$). A 2 (trials) \times 3 (treatments) repeated measures analysis of variance performed on BAT scores resulted in a significant trials main effect, $F(1, 27) = 52.703$, $p < .001$, and a Treatments \times Trials interaction, $F(2, 27) = 3.558$, $p < .05$. Orthogonal comparisons of BAT change scores among the groups showed that the EIP resulted in significantly greater improvement than the ECP, $F(1, 27) = 4.495$, $p < .05$, and that the MIP resulted in significantly greater

improvement than the ECP, $F(1, 27) = 6.062$, $p < .05$. Differences between the MIP and EIP were statistically nonsignificant. The number of posttreatment MIP, EIP, and ECP subjects who actually touched the snake bare-handed (BAT Step 14; Nawas, 1971) were, respectively, 7 , 4 , and 2 .

From the pattern of results that emerged from this study, it is apparent that in comparison to the effect of the imagery procedure used here, therapist empathy had minimal influence on therapy outcome. Neither the inclusion or elimination of therapist behaviors rated as empathetic had a reliable effect on outcome beyond that produced by the imagery procedure.

Using the EIP as a standard of comparison, the MIP group performance indicated that empathy is not a necessary condition for avoidance behavior reduction, whereas the ECP group performance indicated that empathy is not a sufficient condition for the same outcome. These findings are consistent with the interpretations that relationship factors are relatively unimportant in behavioral procedures and that instructed attention shifts toward and away from symptom-related stimuli are important.

References

- Barrett-Lennard, G. T. Dimensions of therapist response as causal factors in therapeutic change. *Psychological Monographs*, 1962, 76(43, Whole No. 562).
- Geer, J. H. The development of a scale to measure fear. *Behaviour Research and Therapy*, 1965, 3, 45-53.
- Nawas, M. M. Standardized scheduled desensitization: Some unstable results and an improved program. *Behaviour Research and Therapy*, 1971, 9, 35-38.
- Wilkins, W., & Domitor, P. J. Role of instructed attention shifts in systematic desensitization. *Proceedings of the 81st Annual Convention of the American Psychological Association*, 1973, 8, 553-554. (Summary)

Received March 4, 1977 ■

Frequent Citations in the *Journal of Consulting and Clinical Psychology* During the 1970s

W. Miles Cox
Indiana University

Authors who have been cited most frequently in the *Journal of Consulting and Clinical Psychology* during the first half-decade of the 1970s are identified. In turn, these authors' publications that have been cited most frequently are indicated. Most prominent are the work of Julian B. Rotter on the locus of control of reinforcement and Truax and Carkhuff's contributions to training and practice in counseling and psychotherapy.

In their recent survey, Koulack and Keselman (1975) found that the *Journal of Consulting and Clinical Psychology* is one of the most prestigious periodicals among clinical psychologists in terms of where they wish to publish and where they expect important psychological material to be found. To determine the influence of certain contributors on publications in the *Journal*, I identified the most frequently cited authors and their most frequently cited publications during the first 5 years of the present decade.

A frequency distribution was constructed to determine the number of times that any given author was cited among the 12,893 reference entries in the 887 articles appearing in Volumes 34-42 (1970-1974). Each author was assigned 1 point for each entry in the reference section of each article in which his or her name appeared. Next, a frequency distribution was constructed to determine the number of references to each publication by the authors who were cited most frequently.

The names of authors who were cited most frequently are shown in Table 1, and their particular publications that were cited most frequently are indicated in Table 2. It should be recognized that frequency of citation might, of course, be a reflection of factors other than an author's influence per se (cf. Cox, 1977). It is noteworthy in this regard that the work of B. J. Winer heads the list of most frequently cited publications as a result of the widespread use of his statistics book (Winer, 1962) rather than his direct contribution to clinical psychology. Nevertheless, from Tables 1 and 2, we can judge what the prominent areas of clinical research have been during the present

Requests for reprints and for an extended report of this study should be sent to W. Miles Cox, who is now at the Psychology Service, Veterans Administration Hospital, 4801 Linwood Boulevard, Kansas City, Missouri 64128.

Table 1
The Most Frequently Cited Authors in the Journal of Consulting and Clinical Psychology, Volumes 34-42, 1970-1974

Author	No. citations	Rank
Rotter, Julian B.	131	1
Truax, Charles B.	105 (1)	2
Gough, Harrison G.	85 (6)	3
Carkhuff, Robert R.	82	4
Cowen, Emory L.	79 (44)	5
Cohen, Jacob	77	6
Lang, Peter J.	76	7
Winer, Ben J.	72	8
Bandura, Albert	70	9
Rogers, Carl R.	68 (1)	10
Zuckerman, Marvin	66 (6)	11
Wolpe, Joseph	65	12

Note. The number of self-references is in parentheses, including self-reference by co-authors.

decade. It should be noted in addition that psychologists who are cited most frequently in the current journal literature are also those who are judged to be scientifically "eminent" on the basis of a variety of other independent criteria (cf. Myers, 1970).

References

- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Bandura, A., & Walters, R. H. *Social learning and personality development*. New York: Holt, Rinehart & Winston, 1963.
- Buss, A. H., & Lang, P. J. Psychological deficit in schizophrenia: I. Affect, reinforcement, and concept attainment. *Journal of Abnormal Psychology*, 1965, 70, 2-24.

Table 2
*The Most Frequently Cited Publications in the
 Journal of Consulting and Clinical Psychology,
 Volumes 34-42, 1970-1974*

Publication	No. citations	Rank
Winer (1962)	72	1
Rotter (1966)	61	2
Truax & Carkhuff (1967)	30	4
Bandura (1969)	24	6
Wolpe (1958)	21	8.5
Rotter (1954)	20	10
Wolpe & Lazarus (1967)	19	11
Gough (1957, 1960, 1964, 1969)	18	12
Gough & Heilbrun (1965)	16	13.5
Lang & Lazovik (1963)	13	16.5
Lang, Lazovik, & Reynolds (1965)	13	16.5
Bandura & Walters (1963)	12	22.5
Buss & Lang (1965)	12	22.5
Cowen, Gardner, & Zax (1967)	12 (4)	22.5
Rogers (1957)	11	28.5
Rogers, Gendlin, Kiesler, & Truax (1967)	11	28.5
Zuckerman, Kolin, Price, & Zoob (1964)	11 (1)	28.5
Lang (1969)	10	32
Lang & Buss (1965)	10	32

Note. The numbers in parentheses refer to self-references.

- Cowen, E. L., Gardner, E. A., & Zax, M. (Eds.) *Emergent approaches to mental health problems*. New York: Appleton-Century-Crofts, 1967.
- Cox, W. M. Some changes in articles published in the *Journal of Applied Psychology* during a twenty-year period. *Journal of Applied Psychology*, 1977, 62, 241-244.
- Gough, H. G. *California Psychological Inventory manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- Gough, H. G. *California Psychological Inventory manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1960.
- Gough, H. G. *California Psychological Inventory manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1964.
- Gough, H. G. *California Psychological Inventory manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1969.

Gough, H. G., & Heilbrun, A. B. *The Adjective Check List manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1965.

Koulack, D., & Keselman, H. J. Ratings of psychology journals by members of the American Psychological Association. *American Psychologist*, 1975, 30, 1049-1053.

Lang, P. J. The mechanics of desensitization and the laboratory study of human fear. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.

Lang, P. J., & Buss, A. H. Psychological deficit in schizophrenia: II. Interference and activation. *Journal of Abnormal Psychology*, 1965, 70, 77-106.

Lang, P. J., & Lazovik, A. D. Experimental desensitization of a phobia. *Journal of Abnormal and Social Psychology*, 1963, 66, 519-525.

Lang, P. J., Lazovik, A. D., & Reynolds, D. J. Desensitization, suggestibility and pseudotherapy. *Journal of Abnormal Psychology*, 1965, 70, 395-402.

Myers, C. R. Journal citations and scientific eminence in contemporary psychology. *American Psychologist*, 1970, 25, 1041-1048.

Rogers, C. F. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 1957, 21, 25-103.

Rogers, C. R., Gendlin, E. T., Kiesler, D. J., & Truax, C. B. (Eds.) *The therapeutic relationship and its impact: A study of psychotherapy with schizophrenics*. Madison: University of Wisconsin Press, 1967.

Rotter, J. B. *Social learning and clinical psychology*. Englewood Cliffs, N.J.: Prentice-Hall, 1954.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 1966, 80(1, Whole No. 609).

Truax, C. B., & Carkhuff, R. R. *Toward effective counseling and psychotherapy: Training and practice*. Chicago: Aldine, 1967.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford, Calif.: Stanford University Press, 1958.

Wolpe, J., & Lazarus, A. A. *Behavior therapy techniques: A guide to the treatment of neuroses*. New York: Pergamon Press, 1967.

Zuckerman, M., Kolin, E. A., Price, L., & Zoob, I. Development of a sensation-seeking scale. *Journal of Consulting Psychology*, 1964, 28, 477-482.

Received March 4, 1977 ■

Predicting Violent Behavior from WAIS Characteristics: A Replication Failure

Lois Shawver and Charles Jew
California Medical Facility, Vacaville, California

Kunce, Ryan, and Eckelman have reported some promising evidence on an index for predicting violent behavior derived from differential Wechsler Adult Intelligence Scale (WAIS) characteristics. The present study is an attempt to replicate their findings. Results of this study were unsuccessful.

Many researchers have tried to use psychological tests to detect people prone to commit violence, but their efforts have generally been disappointing. After reviewing recent negative findings, Kunce, Ryan, and Eckelman (1976) described how they analyzed Wechsler Adult Intelligence Scale (WAIS) profiles of a sample of violent and nonviolent male offenders and factored out new predictors of violent behavior. Their most promising index was the ratio of the subject's Similarity score to the sum of his WAIS subtests $\times 100$. A low Similarities ratio score occurred significantly more often in the violent group in both their original and cross-validation samples. However, since identifying a patient as violent could in itself have devastating consequences, it is imperative that any such findings be thoroughly verified before clinicians put them to use. This should be especially true in small cross-validation samples.

Kahn (1959) advanced a hypothesis similar to the one advanced by Kunce et al. Kahn reasoned, as did Kunce et al., that extreme violence is the result of impulsivity due to poor abstract reasoning ability and should be reflected in lower Similarities scores. Kahn's data were based on Wechsler-Bellevue scores of individuals hospitalized for evaluation of legal insanity. Although Kahn found some elements in the test data supporting his impulsivity hypothesis, his Similarity subtest findings were not as predicted. In contrast to Kunce et al., Kahn found that violent offenders did not have a distinctively lower score on the Similarities subtest.

The present study was an attempt to replicate

the Kunce et al. (1976) findings. Their sample "consisted of white males recently court committed as criminally insane or undergoing pretrial mental examination following arrest for a felony" (p. 42). The sample used here was similar. All subjects were white males diagnosed as psychotic. Subjects were not included when a review of court transcripts left doubt as to whether the crime should be judged as violent or nonviolent.

Kunce et al. reported that violent subjects earned lower Similarities ratio scores ($M = 5.14$) than nonviolent subjects ($M = 9.42$) in a cross-validation sample consisting of 7 violent and 7 nonviolent cases. The present study included 16 subjects in the violent sample and 10 in the nonviolent sample. In the present study, the findings of Kunce and his colleagues were not replicated. Not only were the results not significant, but the direction of the findings were reversed as were the findings reported by Kahn. In this study violent subjects earned higher Similarities ratio scores ($M = 10.25$) than did nonviolent subjects ($M = 9.62$).

The conclusion here is that the Similarities ratio score advanced by Kunce et al. has not yet been sufficiently validated to be used by clinicians to predict violent behavior in specific cases.

References

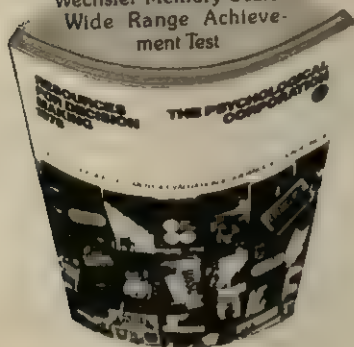
- Kahn, M. W. A comparison of personality, intelligence, and social history of two criminal groups. *Journal of Social Psychology*, 1959, 49, 33-40.
- Kunce, J. T., Ryan, J. J., & Eckelman, C. C. Violent behavior and differential WAIS characteristics. *Journal of Consulting and Clinical Psychology*, 1976, 44, 42-45.

Requests for reprints should be sent to Lois Shawver, Psychological Services, California Medical Facility, Vacaville, California 95688.

Received April 11, 1977

The instruments you need for Clinical Evaluation and Testing.

A-B-C Vision Test • Arthur Point Scale of Performance Tests, Revised Form II • Bayley Scales of Infant Development • Bellak Children's Apperception Test • Bender Visual Motor Gestalt Test • Benton Revised Visual Retention Test • Cattell Infant Intelligence Scale • Columbia Mental Maturity Scale: Third Edition • Cornell Index • Driscoll Play Kit • Dvorine Color Vision Test • Edwards Personal Preference Schedule • Examining for Aphasia (Second Edition) • Farnsworth Dichotomous Test for Color Blindness • Farnsworth-Munsell 100-Hue Test for Color Discrimination • Goldstein-Scheerer Tests for Abstract and Concrete Thinking • Gordon Personal Profile and Gordon Personal Inventory • Harris Tests of Lateral Dominance • Kent Series of Emergency Scales • Holtzman Inkblot Technique • Make A Picture Story (MAPS) • McCarthy Scales of Children's Abilities • Mental Examiners' Handbook • Minnesota Counseling Inventory • Minnesota Multiphasic Personality Inventory • Minnesota Test for Differential Diagnosis of Aphasia • Minnesota Vocational Interest Inventory • Ohio Vocational Interest Survey • Omnibus Personality Inventory • Porteus Mazes • Psychological Screening Inventory • Rorschach Technique • Rotter Incomplete Sentences Blank • San Francisco Vocational Competency Scale • Strong-Campbell Interest Inventory • Survey of Study Habits and Attitudes • System of Multicultural Pluralistic Assessment • Thematic Apperception Test • Thorndike Dimensions of Temperament • Vineland Social Maturity Scale, Revised • Wechsler Intelligence Scales • Wechsler Memory Scale • Wide Range Achievement Test



The Psychological Corporation is a single comprehensive source for evaluation tools that psychologists rely on in clinical settings. Individual intelligence tests. Clinical tests and equipment. Personality and interest inventories. Projective techniques materials. Color vision tests. Tests for the very young and the very old. And every age in between. Tests for the handicapped. The gifted. The culturally-different. Our long list in these areas comprises the most widely-used and respected clinical instruments available in the field of psychology today.

Use the coupon now for more information about these invaluable clinical tools.

THE PSYCHOLOGICAL CORPORATION CLINICAL INSTRUMENTS

Tools for the evaluation of human characteristics, abilities, and behaviors.

The Psychological Corporation

A Subsidiary of Harcourt Brace Jovanovich, Inc.
757 Third Avenue, New York, New York 10017

Yes! Send me a free catalog describing all of The Psychological Corporation's clinical instruments

Name _____
Title _____
Institution _____
Institution Address _____
City _____ State _____ JC&CP 2/78 2¢



THE PSYCHOLOGICAL CORPORATION

A Subsidiary of Harcourt Brace Jovanovich, Inc.
757 Third Avenue, New York, New York 10017



**NATIONAL ACADEMY OF
PROFESSIONAL PSYCHOLOGISTS**
AFFILIATED WITH THE NEW SCHOOL FOR SOCIAL RESEARCH
President, National Academy—**ALFRED J. MARROW, Ph.D.**

Fifth Annual Workshop Schedule

ORGANIZATION BEHAVIOR AND MANAGEMENT

March 31, April 1,2, 1978 Sheraton Ocean Inn, Palm Beach Florida Fee \$300

Leading psychologists will describe the approaches to human resources development used in the country's largest companies. Reports of activities engaged in will include

- 1) organization development,
- 2) behavior modeling, 3) improving the quality of work life,
- 4) assessment techniques,
- 5) management strategies, and more.



WARREN G. BENNIS, Ph.D. Past President, U. of Cincinnati; Chairman, Organizational Studies, M.I.T.



DOUGLAS W. BRAY, Ph.D. Director, Basic Human Resources Research at AT&T since 1956.



DELMAR L. LANDEN, Ph.D. Director, Organizational Research and Development, General Motors Corp.



ALFRED J. MARROW, Ph.D. President, Nat'l. Academy of Professional Psychologists.

ESTABLISHING AND MAINTAINING A PROFESSIONAL PRACTICE

MAY 5, 6, 7, 1978

Assessing readiness for a successful private practice; pre-planning, referrals, budgets, fees, taxes.

A CERTIFICATE OF ATTENDANCE WILL BE ISSUED TO PARTICIPANTS

Registration Form

The National Academy of Professional Psychologists
15 Columbus Circle, New York, N.Y. 10023

Seminars held at the Sheraton Ocean Inn, Palm Beach, Florida,

JCC 2

I would like to attend Seminar(s) _____

Name _____ Highest Degree _____

Home Address _____

City _____ State _____ Zip Code _____

☐ Deposit check enclosed ☐ Payment in full enclosed

\$300 Registration includes tuition and course materials. \$50 Deposit check must accompany registration. Deposit will be refunded if cancelled at least 2 weeks prior to seminar date. Hotel reservations are to be made directly by the participant. Hotel reservations without deposit will be taken on a space available basis only.

These are limited attendance workshops. Make reservations early.

ANNOUNCING!

The Addison-Wesley Series in Clinical and Professional Psychology

edited by Leonard D. Goodstein, *Arizona State University*

To keep you up-to-date, Addison-Wesley offers ten paperback introductions that survey and organize the latest research developments in major sub-areas of clinical psychology. Each is written by an expert or experts in the field and is the ideal middle ground between out-of-date, encyclopedic texts and sketchy journal articles. Whether taken as a whole for introductory level courses or singly for seminars or individual study, this adaptable series can help both students and practitioners learn about and keep in touch with the current state of the field.

Seven titles are being published in 1978:

A Biofeedback Primer

Edward B. Blanchard (SUNY, Albany) and
Leonard H. Epstein (University of Tennessee)
00338, 228 pp, Jan. 1978

Psychotherapy & Process: The Fundamentals of an Existential-Humanistic Approach

James F. T. Bugental (Private Practice, San
Rafael, California; Home Faculty, Humanistic
Psychology Institute) 00333, 192 pp, Feb. 1978

A Primer On Sleep And Dreams

Rosalind Dymond Cartwright (University of
Illinois, Chicago Circle) 00941, 176 pp, Jan. 1978

Psychoanalytic Psychotherapy

edited by George D. Goldman and
Donald S. Milman (Adelphi University)
02513, 192 pp, Jan. 1978

Consulting With Human Service Systems

Leonard D. Goodstein (Arizona State Univer-
sity) 02397, 192 pp, Jan. 1978

Behavior Therapy: A Clinical Introduction

Richard I. Lanyon and Barbara P. Lanyon
(Arizona State University) 04100, 192 pp,
Feb. 1978

Minority Health Issues In Mental Health

William D. Smith, Ann K. Burlew, Myrtis H.
Mosley, and W. Monty Whitney (University of
Cincinnati) 07478, 192 pp, June 1978

Community Psychology

Christopher B. Keys (University of Illinois,
Chicago Circle)

Sex Role Issues In Mental Health

Stephanie A. Shields (University of California
at Davis)

10 Day Free Examination!

Please send me on 10-day examination the books I have checked below. I understand that I may return any of them in ten days and the invoice will be cancelled. Also please send me information on future books in this series as they are published.

- _____ **A Biofeedback Primer** — Blanchard and Epstein (00338) \$4.95
- _____ **Psychotherapy & Process** — Bugental (00333) \$4.95
- _____ **A Primer on Sleep and Dreams** — Cartwright (00941) \$4.95
- _____ **Psychoanalytic Psychotherapy** — Goldman and Milman (02513) \$4.95
- _____ **Consulting with Human Service Systems** — Goodstein (02397) \$4.95
- _____ **Behavior Therapy** — Lanyon and Lanyon (04100) \$4.95

Name _____

Institution/Company _____

Address _____

City _____ State _____ Zip _____

offer valid in U.S. and Canada only

Mail to: Addison-Wesley Publishing Company, Inc.,
Jacob Way, Reading, Mass. 01867

And these three volumes will be published in 1979:

Clinical Human Neuropsychology

Steven G. Goldstein and Susan B. Filskov
(Northside Community Mental Health Center,
Tampa, Florida)

CLINICAL QUANTITATIVE METHODS

By RICHARD W. J. NEUFELD, Ph.D.

CONTENTS: Group Versus Single-Subject Approaches to Obtaining Clinically Relevant Information. Personal Questionnaire Techniques. Statistical Inference in the Single Case. Quantifying Patterns of Symptom Change. Stochastic Process Analysis. Some Methods Facilitating Decision-Making in Clinical Practice. Clinical Use of the Theory of Signal Detectability. Multiple Regression and Discriminant Function Analysis. Canonical Correlation. Introduction to Empirical Classification Approaches. Appendix A: Basic Vector Calculations. Appendix B: Basic Matrix Calculations. Appendix C: Iterative Computation of Latent Roots and Latent Vectors.

1977, 208 pp., 19 illus., \$14.50/£10.30

ISBN: 0-8089-1035-3

THE HUMAN DIMENSION IN PSYCHOANALYTIC PRACTICE

Edited by KENNETH A. FRANK, Ph.D.

CONTENTS: K. A. Frank, Introduction. H. F. Searles, The Development of Mature Hope in the Patient-Therapist Relationship. H. A. Rosenfeld, Personal Experiences in Treating Psychotic Patients. H. Guntrip, My Experience of Analysis with Fairbairn and Winnicott (How Complete A Result Does Psychoanalytic Therapy Achieve?). H. Kelman, My Philosophy of Psychotherapy. H. Bruch, The Tyranny of Fear. R. R. Greenson, That "Impossible" Profession. M. M. R. Khan, From Secretiveness to Shared Living. D. Shainberg, Transforming Transitions in Patients and Therapists. P. Olsen, Recognitions of the Soul. S. Basescu, Anxieties in the Analyst, An Autobiographical Account. A. Ben-Avi, On Becoming an Analyst. E. Singer, The Fiction of Analytic Anonymity.

1977, 224 pp., \$14.75/£10.50 ISBN: 0-8089-1039-6

Send payment with order and save postage plus 50¢ handling charge. Orders under \$15.00 must be accompanied by payment.

Prices are subject to change without notice.

U.S. customers please note: On prepaid orders—payment will be refunded for titles on which shipment is not possible within 120 days.

GRUNE & STRATTON

A Subsidiary of Harcourt Brace Jovanovich, Publishers

111 FIFTH AVENUE, NEW YORK, N.Y. 10003
24-28 OVAL ROAD, LONDON NW1 7DX

Please send me the following:

____ copies, Neufeld: *Clinical Quantitative Methods*

____ copies, Frank: *The Human Dimension in Psychoanalytic Practice*

Check enclosed _____ Bill me _____

NAME _____

ADDRESS _____

CITY/STATE/ZIP _____

New York residents please add sales tax.

Direct all orders to Mr. Paul Negri, Media Dept.
JCCP/2/78

Invited Addresses

Six lectures by eminent behavioral scientists given at the 1976 APA Convention on the theme:

**"Prospects for Control,
Implications
for Freedom."**

Donald O. Hebb
*Science, Psychology,
and Freedom.* Tape 7609.

David Krech —
Prospects for Control.
Tape 7611.

Kenneth B. Clark —
*Problems of Freedom
and Behavior
Modification.* Tape 7612.

B. F. Skinner —
*Human Behavior and
Democracy.* Tape 7610.

Carl R. Rogers —
*Freedom to Be: A
Person-Centered
Approach.* Tape 7613.

David C. McClelland
*Managing Motivation to
Expand Human
Freedom.* Tape 7614.

All tapes are \$5 each,
please use number when
ordering. To order write:

Order Dept. JM.
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036

CHARLES C THOMAS • PUBLISHER

BEHAVIOR THERAPY IN CLINICAL PRACTICE: Decision Making, Procedure and Outcome by Ernest G. Poser, *McGill Univ., Montreal, Quebec, Canada*. Foreword by H. J. Eysenck. Practical applications of social conditioning principles are presented in this text. Divided into four sections, the text explores specific fears, physical expressions of social withdrawal, socially disapproved behavior, and current problems in behavior modification. '77, 204 pp., 4 il., 1 table, \$13.50

MUSIC THERAPY: An Introduction to Therapy and Special Education Through Music (2nd Ptg.) by Donald E. Michel, *Texas Woman's Univ., Denton*. Without extensive use of technical language, this text presents a rationale for the use of music therapy in a wide variety of medical/psychological situations. The history of music therapy; its basic theoretical rationales; and its applications in therapy, rehabilitation, and special education are also discussed. '77, 152 pp., 2 il., \$9.75

DEVELOPMENT AND DISORDERS OF SPEECH IN CHILDHOOD (4th Ptg.) by Isaac W. Karlin; David B. Karlin, *Mount Sinai Hospital, New York City*; and Louise Gurren, *New York Univ., New York City*. In this book the authors describe the normal course and development of speech and language from its origin in infancy through its maturation during adolescence. The anatomical, physiological and psychological aspects of the speech mechanism are presented. The causes of both organic and functional (including emotional) speech and language disorders are also explained. '77, 324 pp., 63 il., 10 tables, \$12.75

THE AMERICAN ALCOHOLIC: The Nature-Nurture Controversy in Alcoholic Research and Therapy (2nd Ptg.) by William Madsen, *Univ. of California, Santa Barbara*. This is a forthright text which calls into question many of the accepted ideas about alcoholism, and it will provide provocative reading for anyone involved in the field. *The Journal of Alcoholism* gave this book a most enthusiastic review: "It was a pleasure to read this thoughtful and scholarly exposition, with its well marshalled facts, and considered opinions, which provide much food for thought and issue many challenges." '77, 272 pp., 4 il., \$11.50

REFLEX EPILEPSY, BEHAVIORAL THERAPY AND CONDITIONAL REFLEXES by Francis M. Forster, *Univ. of Wisconsin, Madison*. The first portion of this singularly comprehensive text deals with the various causes of the condition including visual and auditory stimuli; higher cognitive functions, somatosensory stimulation; and eating and movement. The latter half of the book considers such topics as classification and differential diagnosis, and the methods, principles of application, and role of medication in treatment. '77, 328 pp., 87 il., 22 tables, \$22.75

INDIVIDUALIZED LEARNING PROGRAM FOR THE PROFOUNDLY RETARDED by M. Susan De Vore, *Peoria, Illinois*. This book combines a series of lesson plans with an easy method of collecting data and recording the progress attained by the child. The lesson plans are divided into six sections that individually cover programs for teaching self-care, socialization, communication, cognitive function, gross motor skills and fine motor skills. '77, 256 pp. (9 3/4 x 6 3/4), 35 il., \$11.00, spiral (vinyl)

BASIC APPROACHES TO GROUP PSYCHOTHERAPY AND GROUP COUNSELING (2nd Ed., 2nd Ptg.) edited by George M. Gazda, *Univ. of Georgia, Athens*. (20 Contributors) Experts in the field summarize such techniques as psychodrama, psychoanalysis, existential group psychotherapy, reality therapy, transactional analysis, rational-emotive therapy, conjoint family therapy and gestalt therapy. Group counseling is discussed with respect to the eclectic, behavioral and teleoanalytic positions. '77, 560 pp., 9 il., \$16.75

CLINICAL USE OF PSYCHOTHERAPEUTIC DRUGS (5th Ptg.) by Leo E. Hollister, *Stanford Univ. School of Medicine, Palo Alto, California*. Drugs used in treating emotional disorders of children and adults are covered. Each class of drugs is considered in relation to the chemical and pharmacological differences among its members, pharmacological properties and pharmacokinetics as these relate to their clinical use, specific clinical indications, general principles of use, use in combination with other drugs, side effects, and toxicology. '77, 200 pp., 11 il., 13 tables, \$8.95

Orders with remittance sent, on approval, postpaid

301-327 East Lawrence Avenue • Springfield • Illinois • 62717



Palo Alto School of Professional Psychology

Palo Alto School of Professional Psychology offers a professional graduate program of study leading to the M.A. and Ph.D. in clinical psychology. Committed to the training of effective clinicians, the program provides a broad foundation in psychology and close personal supervision in the training of clinical skills.

PASPP is now accepting applications for admission in April and September, 1978. For catalog and application send one dollar to:

PASPP Admissions Office
467 Hamilton Avenue
Palo Alto, California 94301
(415) 329-0806

NEW

KENT R. JOHNSON
ROBERT S. RUSKIN

Behavioral Instruction: An Evaluative Review

1977 200 pages \$5.00

Provides a comprehensive overview and evaluation of behavioral instruction. Included are sections on the history and development of behavioral instruction; a detailed look at Keller's Personalized System of Instruction, its empirical and theoretical basis and basic principles; other behavioral systems of instruction; management and grading in behavioral instruction courses; comparisons of behavioral and conventional instruction; an extensive component analysis of behavioral instruction; and a discussion of present extensions and future directions of behavioral instruction.

To order this book, please send full payment or institutional purchase order to:



Publications Sales Dept.
American Psychological Association

1200 17th Street, N.W. Washington, D.C. 20036 (202) 833-7600



Keller's Personalized System of Instruction: An Appraisal

Bruce A. Ryan, author; *Foreword* by Fred S. Keller. Offers the reader an objective assessment of the relevance of behavior theory, the role of unit tests, the appropriateness of the student proctor, and the success of self-evaluation in this increasingly popular approach to college-level education. In addition to an analytical review of the current literature, future directions for PSI are proposed and an invaluable "how-to" bibliography is included. 1974. 32 pp. \$2.

To order, send your request accompanied by full payment or institutional purchase order to:



American Psychological Association
Publication Sales Dept.

1200 17th Street, N.W.
Washington, D.C. 20036

Graduate Study in Psychology for 1978-1979

An indispensable
resource book for
prospective
graduate students
and for college
counselors




Graduate Study in Psychology for 1978-1979

This book provides specific information on 500 graduate programs in the United States and Canada: their application procedures, admission requirements, degree requirements, tuition, financial assistance, internships and minority considerations.

It also furnishes general information on how to apply to graduate school and for the first time includes a section of statistical data relevant to trends in graduate training.

600 pages. \$6.

 American Psychological Association
Order Department
1200 17th Street, N.W.
Washington, D.C. 20036

Please send me:

copy(s) of Graduate Study in Psychology
1978-1979 each \$6.

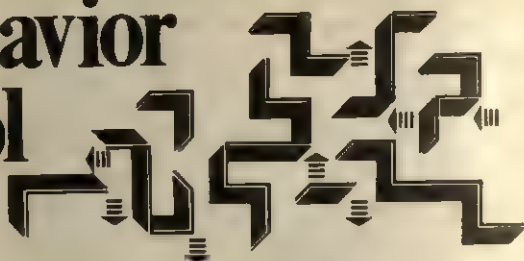
Name _____

Address _____

City _____ State _____ Zip Code _____

Please submit full payment or institutional purchase order.

Master Lectures on Behavior Control



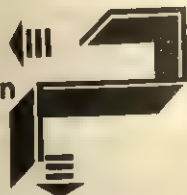
A continuing education program for behavioral scientists. Eight 2-hour lectures presented by eminent psychologists who discuss advances, problems, major concepts, methodologies, and findings in the area of behavior control.

All lectures are in tape cassette format. They may be purchased for \$10 each or the entire 8 lecture set is available at \$75. Payment or an institutional purchase order must accompany all orders.

Other lectures in the series include *Master Lectures on Physiological Psychology*, and *Master Lectures on Developmental Psychology*. Write for a free brochure.



Master Lecture Series/ BC
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036



Irwin Altman, Environmental Control of Behavior and Behavioral Control of the Environment.

Elliot S. Valenstein, The Brain and Behavior Control: Scientific and Social Issues.

William J. McGuire, Persuasion and Social Control.

R. Kirk Schwitzgebel, Legal Issues in Behavior Control.

Nathan H. Azrin, Behavior Modification.

Theodore X. Barber, Hypnosis and Behavior Control.

Daniel X. Freedman, Pharmacology and Behavior Control.

Nicholas Hobbs, Moral and Ethical Issues in Behavior Control.

EIDETICS

a visual system for individual and group psychotherapy by AHSEN

The eidetic image is a bright and lively inner mental picture which elicits a vast spectrum of psychological information. Seen like a movie image, the eidetic causes the person to experience his emotional states with vivid sensations and positive transformation of behavior and consciousness. A normal and fascinating image phenomenon, the eidetic exists in both adults and children, but stays dormant unless helped to become activated through special techniques. Ahseen's acclaimed work on Eidetic Technique, Research, and Therapy Procedures has opened a unique and exciting approach to mind.

SOME VIEWS ON AHSEN'S WORK

- "Unmatched in the clinical literature . . . a methodological advance."
— *The American Journal of Psychiatry*
- "Compared to Akhter Ahseen's penetrating analysis of imagery formation and eidetic processes, all other clinical uses of imagery appear singularly embryonic."
— Arnold A. Lazarus, Ph.D.
- "It is possible that what Ahseen has achieved is a new means of communication, more subtle than verbalization of facts and associative thinking."
— *Behavior Therapy*
- "Adds to currently accepted principles of developmental psychology as exemplified in psychoanalysis and Piaget's work."
— *Pakistan Journal of Psychology*
- "One of the most significant developments yet to emerge in psychotherapy since Freud's psychoanalysis."
— *The Glasgow Journal of Psychology*
- "An exciting and ingenious way of getting at conflict areas."
— *Contemporary Psychology*
- "Really a break-through . . . Its extreme usefulness for divergent schools . . . cannot be denied."
— *Manab Mon, Behaviour and Behaviour Control*
- "Of great value . . . a rapid approach to resolving the patient's problems."
— *Indian Journal of Psychiatry*
- "A very original and very promising approach."
— *Zeitschrift für Psychologie*
- "American Psychiatry is indeed fortunate to have the stimulation and enrichment of Ahseen's highly original and creative presence."
— *Existential Psychiatry*
- "Akhter Ahseen has written an original work which does full credit to Isadore Chien's thesis that philosophy and science can and should be amalgamated to create a systematic view of reality. For that part of the psychological reality with which he deals, Ahseen has given a theoretical framework and a methodology which is one example of how one can unite a metaphysical position with a physical system and put the Cartesian dichotomy aright."
— *International Journal of Clinical and Experimental Hypnosis*



BRANDON HOUSE, Inc. New York, N.Y.

For personal and library orders, write: Brandon House, Inc.,
P.O. Box 240, Dept. 19, Bronx, New York 10471

YES, send me the book(s) checked below. I am enclosing \$. . .
N.Y. residents, please add 8% tax.

- | | |
|---|---------|
| <input type="checkbox"/> PSYCHEYE: Self-Analytic Consciousness | \$9.95 |
| <input type="checkbox"/> BASIC CONCEPTS IN EIDETIC PSYCHOTHERAPY | \$17.50 |
| <input type="checkbox"/> EIDETIC PARENTS TEST AND ANALYSIS:
A Practical Guide to Systematic & Comprehensive Analysis | \$17.50 |

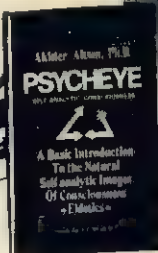
Name _____

Address _____

City _____ State _____ Zip _____

☐ Send bill (Institutions only)

Akhter Ahseen
BASIC
CONCEPTS IN
EIDETIC
PSYCHOTHERAPY



7/12
15

How would you like to use a projective personality test which can be scored quickly and objectively? Whose interpretation is based on careful factor-analytic studies? Which has norms for 11-year-old through adult levels (over 1900 selected Ss)? Uses an interesting and disarming cartoon-like format?

Cassel and Kahn

GROUP PERSONALITY PROJECTIVE TEST

Order Examiner's Set including general manual, scoring keys, 12 test booklets, and 100 answer sheets, enough material for a good-sized testing program (and only \$27.00).

•

For further information, write to

PSYCHOLOGICAL TEST

SPECIALISTS

Box 9229

Missoula, Montana 59807

VOLUME 46, NUMBER 2

APRIL 1978

*Ps.
Dodd &
issue to Mr Liberman.
11 30/11*

JOURNAL OF
CONSULTING
AND
CLINICAL
PSYCHOLOGY

AMERICAN PSYCHOLOGICAL ASSOCIATION

ISSN 0022-006X

The *Journal of Consulting and Clinical Psychology* publishes original contributions on the following topics: (a) the development, validity, and use of techniques of diagnosis and treatment in disordered behavior; (b) studies of populations of clinical interest, such as hospital, prison, rehabilitation, geriatric, and similar samples; (c) cross-cultural and demographic studies of interest for the behavior disorders; (d) studies of personality and of its assessment and development where these have a clear bearing on problems of consulting and clinical psychology; or (e) case studies pertinent to the preceding topics. In order to improve the use of journal resources it has been agreed by the two Editors concerned that the *Journal of Consulting and Clinical Psychology* will no longer consider manuscripts dealing with the etiology or descriptive pathology of abnormal behavior, and the *Journal of Abnormal Psychology* will no longer consider articles dealing with the diagnosis or treatment of abnormal behavior. Articles that appear to have a significant contribution to both of these broad areas may be sent to either journal for editorial decision. Papers of a theoretical nature will also be considered within the space limitations of the journal.

Brendan A. Maher, Editor
Harvard University
Judith P. Worell, Associate Editor
University of Kentucky

Consulting Editors

Thomas M. Achenbach, *National Institute of Mental Health, Bethesda, Maryland*
Kent G. Bailey, *Virginia Commonwealth University*
Barbara B. Brown, *Beverly Hills, California*
James N. Butcher, *University of Minnesota*
Jean Chapman, *University of Wisconsin—Madison*
George M. Chartier, *Arizona State University*
Andrew L. Comrey, *University of California, Los Angeles*
W. Grant Dahlstrom, *University of North Carolina, Chapel Hill*
Anthony Davids, *Brown University*
Gerald C. Davison, *State University of New York at Stony Brook*
Alan S. DeWolfe, *Loyola University of Chicago*
Barbara Dohrenwend, *Graduate School and University Center of the City University of New York*
Richard M. Eisler, *Virginia Polytechnic Institute and State University, Blacksburg, Virginia*
Norman S. Endler, *York University, Downsview, Canada*
Sol Garfield, *Washington University in St. Louis*
Donna M. Gelfand, *University of Utah*
Goldine C. Gleser, *University of Cincinnati Medical Center*
Harrison G. Gough, *University of California, Berkeley*
Jesse G. Harris, Jr., *University of Kentucky*
Mary B. Harris, *University of New Mexico*
Kenneth Heller, *Indiana University*
Frances A. Hill, *University of Montana*
Alexandra G. Kaplan, *University of Massachusetts—Amherst*
Alan S. Kaufman, *University of Georgia*
Alan E. Kazdin, *Pennsylvania State University*
Jeffrey A. Kelly, *University of Mississippi Medical Center*
Barbara K. Keogh, *University of California, Los Angeles*
James C. Kincannon, *University of Minnesota*
Richard I. Lanyon, *Arizona State University*
Herbert M. Lefcourt, *University of Waterloo, Waterloo, Canada*
Peter M. Lewinsohn, *University of Oregon*

Michael J. Mahoney, *Pennsylvania State University*
G. Alan Marlatt, *University of Washington*
Albert Marston, *University of Southern California*
Eric J. Mash, *University of Calgary, Calgary, Canada*
Richard McFall, *University of Wisconsin—Madison*
Martha T. S. Mednick, *University of Connecticut*
Barbara G. Melamed, *University of Florida*
David I. Mostofsky, *Boston University*
Benjamin J. Murawski, *Peter Bent Brigham Hospital, Boston, Massachusetts*
Peter E. Nathan, *Rutgers—The State University*
John M. Neale, *State University of New York at Stony Brook*
K. Daniel O'Leary, *State University of New York at Stony Brook*
Oscar A. Parsons, *University of Oklahoma*
Nolan E. Penn, *University of California, San Diego*
N. Dickon Reppucci, *University of Virginia*
Jerome Resnick, *Temple University*
Leonard G. Rorer, *Miami University*
Gerald M. Rosen, *Providence Family Medical Center, Seattle, Washington*
Robert Rosenthal, *Harvard University*
William Schofield, *University Hospitals, Minneapolis, Minnesota*
Carolyn Simmons, *University of Colorado at Denver*
Bonnie Strickland, *University of Massachusetts—Amherst*
Hans H. Strupp, *Vanderbilt University*
Richard M. Suinn, *Colorado State University*
Richard S. Surwit, *Duke University Medical Center*
Patricia B. Sutker, *Medical University of South Carolina*
Norman F. Watt, *University of Massachusetts—Amherst*
Robert L. Weiss, *University of Oregon*
Joan Welkowitz, *New York University*
G. Terence Wilson, *Rutgers—The State University*
Patricia Wisocki, *University of Massachusetts—Amherst*
Melvin Zax, *University of Rochester*

APA Journal Staff

Anita DeVivo, *Executive Editor*
Deborah S. Felder, *Assistant to the Executive Editor*
Kendall C. Patterson, *Production Supervisor*

Barbara R. Richman, *Production Editor*
Robert J. Hayward, *Advertising Representative*
Juanita Brodie, *Subscription Manager*

Manuscripts: Submit manuscripts in triplicate to the Editor-elect, Sol L. Garfield, Department of Psychology, Washington University, St. Louis, Missouri 63130 according to instructions provided elsewhere in the journal. Authors submitting a manuscript previously considered for publication in another APA journal are invited to inform the Editor, who will then seek to obtain independent reviews, thus avoiding the possibility of repeated reviewing by the same consultant.

Copyright and Permissions: All rights reserved. Written permission must be obtained from the American Psychological Association for copying or reprinting text of more than 500 words, tables, or figures. Permissions are normally granted contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$10 per page, table, or figure of APA material. Permission fees are waived for authors who wish to reproduce their own article. Address requests for permission to the Permissions Office, American Psychological Association, 1200 Seventeenth Street, N.W., Washington, D.C. 20036.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1978: \$32 domestic, \$34 foreign, \$6 single issue. APA member rate: \$16. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Department, APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P.O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Published bimonthly (beginning in February) in one volume per year by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D.C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D.C., and at additional mailing offices.

Diary No. 31
Date 30 April 1978
Tel
March

Journal of Consulting and Clinical Psychology

Copyright © 1978 by the American Psychological Association, Inc.

April 1978 Volume 46, Number 2

Articles

- 207 An Outcome Study of Short-Term Communication Training with Married Couples/
Norman Epstein and Elizabeth Jackson
- 213 Rating Scales for the Identification and Treatment of Hyperkinesis/*Patricia Goldring Zukow, Arnold H. Zukow, and P. M. Bentler*
- 223 Imagery Assessment Through Self-report: What Do Imagery Questionnaires Measure?/*Merrill Hiscock*
- 231 A Model of Psychopathology Based on an Integration of MMPI Actuarial Systems/
Harvey A. Skinner and Douglas N. Jackson
- 239 Multimodal Evaluation of Practice Interactions as Treatment for Social Isolation/
W. Stephen Royce and Hal Arkowitz
- 246 Treatment Outcome and Follow-up Evaluation Based on Client Case Records in a Mental Health Center/*Lynn S. Simons, Teru L. Morton, Terry C. Wade, and Dennis M. McSharry*
- 252 Comparison of Cardiovascular Biofeedback, Neuromuscular Biofeedback, and Meditation in the Treatment of Borderline Essential Hypertension/*Richard S. Surwit, David Shapiro, and Michael I. Good*
- 264 Effects of Transcendental Meditation and Muscle Relaxation on Trait Anxiety, Maladjustment, Locus of Control, and Drug Use/*David C. Zuroff and J. Conrad Schwarz*
- 272 Personality Correlates of Continuation and Outcome in Meditation and Erect Sitting Control Treatments/*Jonathan C. Smith*
- 280 Anxiety: States, Traits—Situations?/*Philip C. Kendall*
- 288 Comparison of Measures of Adaptive Behaviors in Preschool Children/*Linda I. Garrity and Andria B. Servos*
- 294 Effects of Controlled Background Interference on Test Performance by Right and Left Hemiplegics/*Richard E. Nemeo*
- 298 The Enigma of Androgyny: Differential Implications for Males and Females?/
Warren H. Jones, Mary Ellen O'C. Chernovetz, and Robert O. Hansson
- 314 During Training and Posttraining Effects of Live and Taped Extended Progressive Relaxation, Self-relaxation, and Electromyogram Biofeedback/*Irving Beiman, Eileen Israel, and Stephen A. Johnson*
- 322 Correlational and Factor Analysis of the Peabody Individual Achievement Test and the WISC-R/*Richard L. Wikoff*

Comments

- 326 Social Psychological Concepts Applied to Clinical Processes: On the Need for Precision/*Ben Harris and John H. Harvey*
- 329 Imprecision or Dissonance? A Reply to Harris and Harvey/*Robert M. Gordon*
- 331 Selecting a Short Form of the MMPI: Addendum to Faschingbauer/*Norman G. Poythress, Jr.*
- 335 A Note on the MMPI as a Suicide Predictor/*James R. Clopton*
- 337 Response to "A Note on the MMPI as a Suicide Predictor"/*Calista V. Leonard*

(Continued on next page)

Brief Reports

- 339 Characteristics of the Emotional Responsiveness of Sensitizers and Repressors to Social Stimuli/*Glenn A. Miller and William Nuessle*
- 341 Factors Influencing Lower-Class Black Patients Remaining in Treatment/*Anthony Vail*
- 342 Effects of Explanation and Information Feedback on the Illusory Correlation Phenomenon/*Ronald W. Waller and Stuart M. Keeley*
- 344 The 4-3 MMPI Profile Type: A Failure to Replicate/*Jeffrey A. Buck and John R. Graham*
- 345 Congruence of Parental Perception, Marital Satisfaction, and Child Adjustment/*Lucy Rau Ferguson and Deborah R. Allen*
- 347 A Factor Analysis of Assertive Behaviors/*Joseph S. Pachman, David W. Foy, Frank Massey, and Richard M. Eisler*
- 348 Life Change, the Sensation Seeking Motive, and Psychological Distress/*Ronald E. Smith, James H. Johnson, and Irwin G. Sarason*
- 350 Relation of Patient Attributes to Perceptions of the Treatment Environment/*Rudolf H. Moos and Evelyn Bromet*
- 352 Effect of Psychotomimetics (LSD and Dextroamphetamine) on the Use of Primary- and Secondary-Process Language/*Michael Natale, Charles C. Dahlberg, and Joseph Jaffe*
- 354 Short Forms of the MMPI with Back Pain Patients/*Judith Turner and Charles McCreary*
- 356 Practicing A and B Psychotherapists' Responses to Schizoid and Neurotic Patient Prototypes/*William B. Goodwin, Donald M. Quinlan, and Jesse D. Geller*
- 358 Comparisons of Rating Scales of Child Psychopathology in Clinic and Nonclinic Samples/*Susan B. Campbell and Yvonne Steinert*
- 360 Buss-Durkee Assessment and Validation with Violent Versus Nonviolent Chronic Alcohol Abusers/*Gisele J. Renson, John E. Adams, and Jared R. Tinklenberg*
- 362 On the Word Preferences of Suicidal Versus Nonsuicidal College Students/*Steven Thurber and David P. Torbet*
- 364 Role of Locus of Control in Frustration-produced Aggression/*Kiran Bhatia and Sanford Golin*
- 366 Personality and Social Factors in Adolescent Marijuana Use: A Path-Analytic Study/*Paul M. Kohn and Helen M. Annis*
- 368 Weight Loss and Behavior Change 1 Year after Behavioral Treatment for Obesity/*Robert W. Jeffery, Michael Vender, and Rena R. Wing*
- 370 Taylor's MMPI Correction Factor for Spinal Cord Injury: Empirical Endorsement/*Philip C. Kendall, Jack Edinger, and Carole Eberly*
- 372 Meanings that Professionals Attach to Labels for Children/*Charles F. Carroll and N. Dickon Reppucci*
- 375 Effects of Anxiety and Sex on Neuropsychological Tests/*Glen D. King, H. Julia Hannay, Bruce J. Masek, and Joan W. Burns*
- 377 Diversity Among Lower-Class Therapy Clients: A Comparison of Class IV and Class V Psychotherapy Recipients/*Michael T. Nietzel, Matthew G. Hile, and Charles Y. Kondo*
- 379 Normative Data on Locus of Control in Male Alcoholics Undergoing Treatment/*Frederick W. Obitz*
- 381 Distinguishing Learning-Disabled and Emotionally Disturbed Children on the WISC-R/*Raymond S. Dean*
- 383 Application of Electromyographic Biofeedback to the Relaxation Training of Schizophrenic, Neurotic, and Tension Headache Patients/*Frank X. Acosta, Joe Yamamoto, and Stuart A. Wilcox*

Other

- 212 Garfield Appointed Editor, 1979-1984
- 230 Instructions to Authors

An Outcome Study of Short-Term Communication Training with Married Couples

Norman Epstein and Elizabeth Jackson
State University of New York at Buffalo

Communication training, interaction insight training, and no treatment were compared for changes in marital verbal interaction and spouses' ratings of each other on the Barrett-Lennard Relationship Inventory. Fifteen couples were randomly assigned to the three groups. Both treatments involved five 1½-hour group sessions over 3 weeks, led by the same male and female cotrainers. During assessments, discussions of individual couples were tape recorded and coded. The pretest-posttest interval for waiting list controls was equal to that for the treatment groups. Communication training produced a significant increase in assertive requests, compared to insight treatment and no treatment. Both treatments reduced disagreement significantly. Communication training produced a greater decrease in attacks and a greater increase in spouse-rated empathy than the control condition, but insight training and no treatment did not differ on these variables. Generally, communication training led to more extensive changes in spouses' verbal behavior and perceptions of marital communication than insight training. Further research is suggested to test the limits of this intervention for modifying various classes of maladaptive marital interaction.

In recent years, treatments of marital discord frequently have been designed to foster clear communication between intimates, with the expectation that such behavioral changes would be associated with increased marital satisfaction (e.g., Bolte, 1970; Ely, Guernsey, & Stover, 1973; Lederer & Jackson, 1968; Wells, Figurel, & McNamee, 1975). Raush, Barry, Hertel, and Swain (1974) noted that conflict is inevitable when partners in a close relationship seek to satisfy their varied needs, and they argue that clear communication is a prerequisite for conflict resolution. Although poor communication may be a consequence of marital conflict, the lack of information exchange may itself impede resolution of differences and interpersonal tensions (Bardill, 1966; Raush et al., 1974). It is likely that

poor communication and marital dissatisfaction mutually reinforce each other.

Empirical investigations of the relationship between communication clarity and marital satisfaction provide evidence that their association is quite strong. Relying on self-report measures, Navran (1967) and Murphy and Mendelson (1973) found high positive correlations between spouses' scores on marital satisfaction inventories and their scores on questionnaires assessing openness of marital communication. In Raush et al.'s (1974) study, spouses' use of rejection and coercion, coded from actual interactions, was significantly related to poor resolution of conflict. Relationships characterized by avoidance had a static quality in which little problem solving took place. Epstein and Jackson (Note 1) examined the relationships between spouses' ratings of empathy, congruence, and unconditional positive regard received from each other and categories of their verbal interaction coded from tapes by independent raters. Perceived empathy was related inversely to frequencies of self-justification and disagreement

This article is based on a paper presented at the meeting of the Eastern Psychological Association, New York, April 1976.

Requests for reprints should be sent to Norman Epstein, Department of Psychology, State University of New York at Buffalo, Buffalo, New York 14226.

Copyright 1978 by the American Psychological Association, Inc. All rights of reproduction in any form reserved.

by the spouse and related positively to the spouse's use of statements that revealed feelings, offered support, or conveyed agreement. Ratings of the spouse's congruence were related inversely to the frequency of self-justification. These findings suggest that patterns of clear, supportive communication are associated with components of marital satisfaction.

Two basic patterns of unclear communication commonly observed in disturbed marriages are avoidance and active escalation of conflict. Olson (1972) and Raush et al. (1974) suggested that spouses who feel threatened by potential confrontation on relationship issues often withdraw from direct communication. They rely on evasive messages such as denials, self-justification, and disqualification¹ rather than stating their opinions and feelings clearly. On the other hand, conflicts may be escalated directly through issue expansion (e.g., "You always _____"), disparagement, and other personal attacks (Raush et al., 1974). Both avoidance and escalation of attack decrease the exchange of information necessary for constructive problem solving.

Avoidance and attack closely parallel the two modes of behavior described by Alberti and Emmons (1974) as characteristic of people who have deficits in assertive behavior: nonassertion (withdrawal) and aggression. They outline procedures designed to improve an individual's communication skills by substituting specific assertive messages for evasive and/or attacking messages. Although group assertiveness training has been shown to increase clarity and directness of subjects' communication in role playing situations, Alberti and Emmons question whether the newly learned behaviors will generalize to interactions with significant others outside the training group. Studies by Eisler and Hersen (1973) and Lehman-Olson (1976) provide some encouraging evidence for the effectiveness of assertiveness training with individuals, focusing on their marital relationships. However, both Alberti and Emmons and Lehman-Olson suggested that assertiveness training with married persons would be more effective if spouses participated together. The present study draws on tech-

niques of assertion training for in vivo modification of unclear communication between spouses.

In the present study, a group treatment designed to increase clarity and assertiveness of communication by means of active practice (e.g., role playing) was administered to couples who complained of difficulties in communicating. Its effects on both verbal behavior and spouses' self-reported perceptions of each other's empathy, congruence, and positive regard were compared to those found with an alternate treatment focusing on insight into behavioral interaction patterns and with a no-treatment control group. The "interaction insight" treatment fostered spouses' understanding of repetitive maladaptive behavioral patterns (e.g., interruptions) in their relationship, without the repeated practice of assertive messages incorporated in the communication treatment. Inclusion of this alternate treatment group allowed a test of the hypothesis that active practice of more effective communication patterns will produce not only more assertive behaviors but also a greater increase in spouse-perceived openness of communication than that resulting from insight training alone.

The specific hypotheses were that in a pretest-posttest design, (a) subjects who received communication training would exhibit a greater increase in specific assertive messages than subjects in the interaction insight and no-treatment control groups; (b) subjects receiving communication training would exhibit a greater decrease in nonassertive messages (e.g., disqualification, self-justification) and aggressive messages (e.g., attack) than subjects in the other groups; (c) subjects who received communication training would report greater increases in their spouses' congruence (open, direct communication) than subjects in the other groups;

¹ A disqualification is a message that invalidates another message sent previously or concurrently. Disqualifying one's own message allows one to express an idea or feeling but not take responsibility for it. Disqualifying another person's message communicates that one does not acknowledge the validity or even the existence of the other's ideas and feelings (Olson, 1972; Watzlawick, Beavin, & Jackson, 1967).

(d) subjects in both the communication and interaction insight groups would report increases in empathy and positive regard received from their spouses, with the former treatment producing greater increases on these dimensions; and (e) improved conflict resolution, evidenced by fewer disagreements, would result from both treatments, with communication training producing the stronger effect.

Method

Subjects

Couples who responded to public media announcements of a research project involving free workshops for "marital communication problems" initially were screened by phone in order to select subjects who were married at least 2 years, had completed high school, were not presently in counseling, were not judged to be in psychological crisis, and did not exhibit severe psychopathology. These criteria were used to insure a moderate degree of homogeneity in the sample. Appropriate referrals were made for callers who did not meet the above criteria.

A pretreatment interview with each couple served as a second screening. Three interviewed couples were referred for more extensive marital therapy based on the experimenters' judgments that their problems were more severe than a communication disorder. The remaining 16 couples who were judged appropriate for the workshops then were randomly assigned to the communication training group (6 couples), the interaction insight group (5 couples), and the no-treatment "waiting list" group (5 couples). During the course of treatment, 1 couple dropped out of the insight group. Subjects placed on the waiting list received treatment at the conclusion of the study.

Materials and Procedure

During pretreatment and posttreatment assessments for individual couples, each subject completed the Barrett-Lennard Relationship Inventory (Barrett-Lennard, 1962), indicating the degree of empathy, congruence, and unconditional positive regard generally received from the spouse. This scale has been used in previous studies (e.g., Quick & Jacob, 1973; Wells, Figurel, & McNamee, 1975) to assess marital satisfaction.

Next, the couple was instructed to discuss their communication problems for 15 minutes with a male and a female interviewer present. The interviewers did not intervene in the couple's discussion after giving the following instructions:

In order for us to get an idea about how the two of you talk with each other, we would like

to spend 15 minutes giving each of you an opportunity to talk about what communication is like between you. What seems to happen when you want to talk to each other about important matters? If there is a problem with your communication process, what is it, and what seems to be its source? Please direct your comments to each other, not to us, and we will listen.

Following this warm-up period, the couple was asked to continue discussing their communication pattern for 10 minutes without the interviewers present. Interviews were tape recorded with each couple's written permission.

Subjects in the two treatment groups met for five 1½-hour sessions over a 3-week period. Both groups were led by the same male and female cotrainers, who were experienced in both modes of treatment and were unaware of the experimental hypotheses. Although the treatment formats were different for the two groups, the major topic discussed at each session (e.g., expression of caring, cooperation, problem solving) was the same. During each session, subjects in both treatments formed smaller groups of two or three couples to perform specific exercises. At the beginning and end of each session, the total group membership met to discuss common communication issues.

The communication treatment emphasized the practice of specific assertive requests, opinions, and statements of feeling. During the first session, couples were presented a set of guidelines for clear communication. Alberti and Emmons' (1974) arguments regarding "assertive rights" in interpersonal relationships were discussed. All sessions included modeling of assertive communication by cotrainers, practice by each couple with issues they chose from their own relationships, and behavioral feedback to each subject from the trainers and other group members. Instances of unclear and nonassertive communication were pointed out, and alternate ways of expressing the intended message were presented for the individual to role play. Couples increasingly coached each other on expressing messages precisely and directly. The cotrainers intervened actively and reiterated the communications guidelines frequently.

The insight treatment focused on the particular interaction patterns of each couple that confused or frustrated the spouses. The cotrainers instructed subjects in the observation of verbal and nonverbal messages that exacerbate conflict. A major goal was to increase each subject's awareness of the impact his/her behavior had on the spouse's feelings and behavior. Each couple received extensive feedback regarding interaction patterns from the cotrainers and other couples. Delineation and practice of alternate modes of clear communication was minimal.

Each couple's posttest interview was held within a week of the last treatment session and was identical in procedure to the pretest. The pretest-posttest interval for the control group was equal to that for the treatment groups. All couples were given an opportunity for further treatment.

Results

Eleven categories of each subject's taped verbal behavior were scored by two independent coders from a 5-minute segment of each couple's discussion while alone. Based on an a priori decision to allow a short time for each couple to adjust to the interviewers' departure, the coded segment always was the 3rd through the 7th minute of discussion. The coding categories and their respective interrater reliabilities, calculated as phi coefficients (Herbert, Note 2), were assertive request, 1.00; self-revelation, .85; specific statement, .86; general statement, .83; asking for feedback, 1.00; disqualification, .90; agreement, .98; disagreement, .80; support, 1.00; attack, .84; and self-justification, .84.

Pretest, posttest, and change (posttest minus pretest) scores were computed for each subject on the proportion of total speech acts coded into each of the 11 verbal behavior categories and ratings of spouse's empathy, congruence, and positive regard. Since spouses were assessed as couples, their scores on each dependent measure were not considered to be independent, and the analyses were conducted with couples' scores computed as the mean score for each husband and wife. Therefore, the *ns* for analyses were 6, 4, and 5 couples in the communication, insight, and control groups, respectively.

Analyses of variance for pretest scores indicated no significant initial differences among the three groups on the dependent variables. The fact that only 1 of the 14 pretest *F* values reached the level of $p = .10$, which would be expected by chance alone, and only 2 reached $p < .20$, served as evidence that the random assignment of couples to groups produced adequate matching.

Further one-way analyses of variance and planned (one-tailed) *t* tests were used to compare change scores for the groups. The potential problem of highly correlated dependent measures, which would result in redundant tests in the major analyses, was assessed by computing the intercorrelations of all the dependent variables. The mean absolute value of these 91 correlations, calculated with z_r transformations (Edwards, 1976), was .22 (*ns*) and represents a mean

of less than 5% shared variance. Although 3 of the 7 correlations exceeding .40 (16% overlap) were those among the three Relationship Inventory subscales, the important conceptual distinctions inherent in the subscales and the differential predictions made for them in this study were considered to be sufficient grounds for analyzing them separately.

Of the one-way analyses of variance for the three subject groups, the strongest effects were found for the following dependent variables: assertive request, $F(2, 12) = 27.65$, $p < .001$; disagreement, $F(2, 12) = 3.42$, $p = .067$; attack, $F(2, 12) = 2.08$, $p = .168$; spouse-rated empathy, $F(2, 12) = 1.75$, $p = .215$; asking for feedback, $F(2, 12) = 1.69$, $p = .225$; and self-justification, $F(2, 12) = 1.37$, $p = .290$. No other variable produced an *F* value exceeding 1.00.

Planned *t* tests indicated a significant increase in assertive requests for the communication group, relative to the insight group, $t(12) = 6.51$, $p < .001$, and relative to the control group, $t(12) = 6.03$, $p < .001$. Change in assertive requests did not differ for the insight and control groups, $t(12) = -.82$. Significantly greater reduction in percentage of disagreements was found for the communication group, $t(12) = -1.84$, $p < .05$, and the insight group, $t(12) = -2.52$, $p < .05$, relative to the control group. Contrary to the hypothesis, the decrease in disagreement was not significantly different for the insight and communication groups, $t(12) = .89$.

As predicted, communication training produced a significantly greater decrease in attacks than the no-treatment control, $t(12) = -2.03$, $p < .05$. Insight training produced an intermediate decrease in attacks that did not differ significantly from the other groups. There were nonsignificant trends toward decrease in use of self-justification in the communication group, relative to the control group, $t(12) = -1.41$, $p < .10$, and in the insight group, relative to the control group, $t(12) = -1.46$, $p < .10$. Similarly, there were nonsignificant trends toward an increase in asking for feedback in the communication group, relative to the control group, $t(12) = 1.42$, $p < .10$, and in the insight group, rela-

tive to the control group, $t(12) = 1.72$, $p < .10$. No other significant differences were found among the three groups for the remaining verbal coding categories.

Analyses of subjects' change scores on the Relationship Inventory indicated a significant increase in spouse-perceived empathy in the communication group, relative to the control group, $t(12) = 1.79$, $p < .05$. The trend toward greater increase in empathy for the communication group than for the insight group was not significant, $t(12) = 1.25$. The trend toward greatest increase in spouse-rated congruence in the communication group also was not significant.

Table 1 presents the pretest, posttest, and change-score group means for the significant effects.

Discussion

The significant decrease in disagreements of subjects in the two treatment groups suggests that they were moderately successful in reducing overt conflict in this sample of married couples. Although communication training did not produce a greater decrease in overt disagreement than interaction insight

training, the communication training was more effective in reducing attack behaviors. The somewhat smaller decrease in statements of disagreement in the communication group might be due to the fact that the assertiveness training encouraged direct expression of feelings and opinions, including disagreement. An unanswered question is whether or not training in the assertive expression of disagreement along with decreased attacking will lead to greater conflict resolution than insight training in the long run.

The significant increases in assertive requests and spouse-perceived empathy in the communication group indicate that a short-term structured intervention involving both partners can have a measurable impact on both overt interactions and spouses' experiences of each other's communication. The overall findings provide evidence that communication training produces greater changes in some categories of subjects' behavior than interaction insight training. The lack of significant change in other categories of verbal behavior (e.g., disqualifications, self-revelation) and in spouse-perceived congruence and unconditional positive regard suggests that the effectiveness of this communication training for couples should be tested further. It is important to determine whether treatment of longer duration would produce comparable change across the various communication categories or whether certain classes of behavior are particularly resistant to change with this intervention. Since the impact of communication training on spouses' perceptions of change was less extensive, it appears that it may be easier to implement behavioral changes than attitudinal changes in close interpersonal relationships. A question for future research is whether spouses' perceptions of behavioral change will follow in time.

Reference Notes

1. Epstein, N., & Jackson, E. *Marital communication disorders: Behavioral patterns and spouse perceptions*. Unpublished manuscript, State University of New York at Buffalo, 1976.
2. Herbert, E. W. *Computation guide for interobserver reliabilities*. Unpublished manuscript, Bureau of Educational Research, University of Utah, 1973.

Table 1
Pretest, Posttest, and Change Score Group Means for Significant Effects

Group and score	Variable			
	Assertive request	Attack	Disagreement	Spouse-rated empathy
Communication ^a				
Pretest	.00	.29	.15	-8.75
Posttest	.08	.18	.09	-.25
Change	.08	-.11	-.06	8.50
Insight ^b				
Pretest	.01	.28	.24	-7.38
Posttest	.00	.23	.12	-6.26
Change	-.01	-.05	-.12	1.12
Control ^c				
Pretest	.01	.15	.10	3.40
Posttest	.01	.19	.14	2.00
Change	.00	.04	.04	-1.40

^a $n = 6$.

^b $n = 4$.

^c $n = 5$.

References

- Alberti, R. E., & Emmons, M. L. *Your perfect right* (2nd ed.). San Luis Obispo, Calif.: Impact, 1974.
- Bardill, D. R. A relationship-focused approach to marital problems. *Social Work*, 1966, 2, 70-77.
- Barrett-Lennard, G. T. Dimensions of therapist response as causal factors in therapeutic change. *Psychological Monographs*, 1962, 76(43, Whole No. 562).
- Bolte, G. L. A communications approach to marital counseling. *The Family Coordinator*, 1970, 19, 32-40.
- Edwards, A. L. *An introduction to linear regression and correlation*. San Francisco: Freeman, 1976.
- Eisler, R. M., & Hersen, M. Behavioral techniques in family oriented crisis intervention. *Archives of General Psychiatry*, 1973, 28, 111-116.
- Ely, A. L., Guernsey, B. G., & Stover, L. Efficacy of the training phase of conjugal therapy. *Psychotherapy: Theory, Research and Practice*, 1973, 10, 201-207.
- Lederer, W. J., & Jackson, D. D. *The mirages of marriage*. New York: Norton, 1968.
- Lehman-Olson, D. Assertiveness training: Theoretical and clinical implications. In D. H. L. Olson (Ed.), *Treating relationships*. Lake Mills, Iowa: Graphic, 1976.
- Murphy, D. C., & Mendelson, L. A. Communication and adjustment in marriage: Investigating the relationship. *Family Process*, 1973, 12, 317-326.
- Navran, L. Communication and adjustment in marriage. *Family Process*, 1967, 6, 173-184.
- Olson, D. H. Empirically unbinding the double-bind: Review of research and conceptual reformulations. *Family Process*, 1972, 11, 69-91.
- Quick, E., & Jacob, T. Marital disturbance in relation to role theory and relationship theory. *Journal of Abnormal Psychology*, 1973, 82, 309-316.
- Raush, H. L., Barry, W. A., Hertel, R. K., & Swain, M. A. *Communication, conflict, and marriage*. San Francisco: Jossey-Bass, 1974.
- Watzlawick, P., Beavin, J. H., & Jackson, D. D. *Pragmatics of human communication*. New York: Norton, 1967.
- Wells, R. A., Figurel, J. A., & McNamee, P. Group facilitative training with conflicted marital couples. In A. S. Gurman & D. G. Rice (Eds.), *Couples in conflict*. New York: Aronson, 1975.

Received June 14, 1976 ■

Garfield Appointed Editor, 1979-1984

The Publications and Communications Board of the American Psychological Association announces the appointment of Sol L. Garfield as Editor of the *Journal of Consulting and Clinical Psychology* for the years 1979-1984. As of January 1, 1978, manuscripts should be directed to the Editor-elect:

Sol L. Garfield
Department of Psychology
Washington University
St. Louis, Missouri 63130

Rating Scales for the Identification and Treatment of Hyperkinesis

Patricia Goldring Zukow, Arnold H. Zukow, and P. M. Bentler
University of California, Los Angeles

In response to the need for a simple instrument to aid the psychologist, educator, or physician in the identification and treatment of hyperkinesis, parent and teacher rating scales were developed. Multivariate analysis of the parent ratings of clinical and control subjects indicated three factors: Factor 1, Excitability; Factor 2, Motor Coordination; and Factor 3, Directed Attention. A factor analysis of the teacher form yielded two similar factors, Attention/Excitability and Motor Coordination. Analysis of variance of each factor score revealed highly significant differences between clinical and control subjects. Cutoff scores were developed to aid in diagnostic decision making. These scores correctly identified a large percentage of clinical and control subjects.

Many concerned professionals have expressed well-founded reservations about the use of medication (Cantwell, 1975; Fish, 1975; Myers & Pless, 1976; Rie, 1975; Rie, Rie, Stewart, & Ambuel, 1976; Sroufe & Stewart, 1973; Grossman, Note 1) and the effectiveness of various therapeutic approaches (Gittelman-Klein et al., 1976; Keogh & Margolis, 1976) in the management of hyperactive children. Several complete reviews and analyses of both the positive effects as well as the drawbacks of drug therapy have recently appeared (Rie et al., 1976; Whalen & Henker, 1976). In spite of the fortunate fact that medication can yield behavioral improvement in a large proportion of cases of hyperkinesis (Conners, 1972; Conners, Rothschild, Eisenberg, Schwartz, & Robinson, 1969; Conners, Taylor, Meo, Kurtz, & Fournier, 1972; Sleator & von Neumann, 1974) or that behavior modification techniques affect short-term or situational improvement in academic performance (Gittelman-Klein et al., 1976; O'Leary, Pelham, Rosenbaum, & Price,

1976; Varni, 1976), the specific biological and behavioral bases and diagnostic criteria for hyperkinesis remain uncertain. There is common agreement that hyperkinesis contains elements of high activity level associated with learning and behavioral disorders (Laufer & Denhoff, 1957; Millichap, 1968; Stewart, Pitts, Craig, & Dieruf, 1966; Werry, 1968), but the practical issue of identification remains unresolved (Rie, 1975; Grossman, Note 1).

Many physical and behavioral disorders can be reliably classified on the basis of first-hand observations by a single professional. However, the identification of hyperkinesis must involve the active cooperation of both parents and teachers. One study reported, for example, that only 10 of 46 subjects could have been correctly identified on the basis of a single examination conducted by one individual (Sleator & von Neumann, 1974). As Rie (1975) has noted, satisfactory assessment of a child's behavior must be based on evaluations by several persons well acquainted with many facets of the child's everyday life. It follows that the correctness of classification, consequent assignment of treatment condition or therapy, and assessment of behavioral changes will depend heavily on the adequacy of behavioral measurement provided by teachers and parents.

This study was supported by U.S. Department of Health, Education, and Welfare National Institute on Drug Abuse Grant DA1070.

Requests for reprints should be sent to Patricia G. Zukow, Department of Psychology, University of California, Los Angeles, California 90024.

Teachers are acquainted with a broad range of children engaged in a great number of school activities and can thus compare children's behavior to standards for their age and community; parents, of course, can provide observations from their daily contact with their own children and others in a variety of naturalistic settings.

There is as yet little consensus regarding the perceived behavioral bases of the hyperkinetic syndrome. Several comprehensive behavior rating scales designed to be used by parents and teachers in evaluating problem behavior succeed in differentiating between hyperkinetic and normal children (Blunden, Spring, & Greenberg, 1974; Conners, 1969, 1970; Kupietz, Bialer, & Winsbert, 1972). However, these instruments do not specify the facets of behavior involved in hyperkinesis. So far, the instruments designed for the specific purpose of identifying hyperkinesis have been inadequately researched (Davids, 1971) or they evaluate behavior too specific in nature to capture the clinical complexity of hyperkinetic syndrome (Werry & Sprague, 1970). Blunden et al. (1974) identified a single hyperactivity factor in their teacher rating data of normal subjects. Aspects of hyperactivity included restlessness, impulsiveness, and low perseverance, in addition to the concepts of distractibility and low concentration. Thus in the Blunden et al. study, the presence of one aspect of hyperactivity implied the presence of the others. Conners (1969), in contrast, has found that distractibility and concentration difficulties, along with poor coordination, represented a daydreaming/inattentiveness factor that was clearly distinct from hyperactivity in teacher rating scales. According to teachers a hyperactive child might, or might not, also be shown to have coordination and concentration problems. Finally, Conners (1970) demonstrated that parents who rated behavioral symptoms clustered their symptoms into several distinct dimensions, but that it was primarily one dimension—aggressive conduct disorder—that proved to differentiate normal versus clinic patients and neurotic versus hyperkinetic children. Although this factor makes sense as an identifier of hyperkinesis

because it includes behaviors such as restlessness and temper, the dimension seems to include irrelevant behaviors such as having problems with peers and siblings and excludes such relevant behaviors as poor coordination and inattentiveness that may lead to problems in school. It is apparent that even though some differential diagnosis is possible with existing instruments, the behavioral basis for the disorder remains uncertain.

A recent study by Langhorne, Loney, Paternite, and Bechtoldt (1976) analyzed certain chart, teacher, parent, and psychiatrist ratings of a variety of variables considered to be relevant to the diagnosis of hyperkinesis. Using factor analysis they found that chart raters and psychiatrists were using the same basis for judgments of hyperactivity. Chart raters and teachers tended to assess excitability, but only teachers agreed regarding information about attention difficulties. Although Langhorne et al. interpreted their results as reflecting source of data rather than intrinsic characteristics of hyperkinesis, it is difficult to evaluate their results. All their analyses were based on only a very small, inadequately rationalized subset of variables, which were also subjected to unnecessary transformation prior to analysis. It is clear that additional studies of this issue are called for.

The current research is addressed to two interrelated goals. First, it aims at clarifying the behavioral bases of hyperkinesis as seen in parent and teacher reports. Second, it aims to provide a measure of hyperkinetic behavior that is relevant to the objective classification and treatment of hyperkinesis and to the assessment of significant behavioral changes.

Parent Form

Method

Initial item pool. Since it seemed possible that hyperkinesis as perceived by the parent might involve several interrelated, but distinct, aspects of behavior, items were written by the second author to cover these areas. The first cluster of items was intended to assess "classical" hyperkinetic behavior, including such behaviors as being quick-tempered and explosive. Items and descriptions found in

other scales and observed in clinical rounds, conferences, and journals were modified and adopted. The second group of items was directed toward assessing motor coordination, such as the ability to button without difficulty. These items were included to clarify the observation that poor coordination seems to accompany daydreaming and inattentive behavior more closely than hyperactivity per se (Conners, 1969). The third type of item was concerned with the inability to sustain continuous participation in, and completion of, structured activities related to learning and school problems, as perceived by the parent. This includes behavior such as daydreaming, which had been found in previous work to be possibly distinct from the first item type but which was clearly implicated in the hyperkinetic syndrome (Conners, 1969). A total of 28 items were written. Instructions required the parent, usually the mother, to "circle the answer that most applies to your child." Possible responses were yes or no, although one item was a three-category item collapsed to yes/no for analysis.¹

Subjects. The questionnaire was administered to mothers of children attending cooperating public elementary schools and private preschools and, after initial screening for hyperkinetic behavior problems, to mothers of a subset of the children referred to a private pediatric practice. At the time of evaluation the children were not receiving medication for hyperkinetic behavior. Mothers of the public school children received the questionnaire from teachers who were selected by the principal or director of the schools. Completed forms were returned to the teachers. The schools were chosen to be representative of the socioeconomic class of the children predominantly in this private practice, that is, skewed slightly toward upper-middle class. Various private physicians, mental health practitioners, school personnel, and concerned nonprofessionals referred children to a private pediatric practice for evaluation of possible hyperkinetic disorder. The judgments of the examining pediatrician, the second author, were based on physical examination, medical and family histories, plus behavioral evidence reported by parents and teachers. Inclusion in the clinical group depended on consistent evidence of behavioral disturbance at home and at school along with a persistent learning disorder associated with an inappropriately high activity level. The clinical group children were judged not to possess an overt neurological syndrome, deafness, visual disability, or global mental retardation. Other rare but possible causes of hyperactivity such as hypoglycemia or hyperthyroidism were excluded by appropriate laboratory tests.

Data were available from mothers of 136 control and 78 clinical children, out of 278 questionnaires distributed. The control sample was composed of a nearly equal number of males as females. All ratings of male subjects were retained, and a subset of female subjects was randomly selected from the control sample to best approximate the ratio of males to females in the clinical group (3:1). Final

sample sizes at 2½-5½ years of age were 63 control and 27 clinical subjects, and at ages 5½-11, 19 control and 51 clinical children.

Method of analysis. There were three separate steps to the analysis. First, items were grouped into three clusters on the basis of their statistical consistency with other items. An overall score was then obtained for each group of items, yielding three combined scores and a fourth total score across all items. Finally, analyses of variance were performed to evaluate whether the four combined scores yielded significant differences among clinical, control, and age-varying children.

Only 26 out of the 28 items on the questionnaire proved analyzable due to an error in data collection. To determine the interrelations among these items, in the total sample of subjects, monotonicity analysis was deemed appropriate, since the item responses were binary. This procedure determines latent dimensions of variation for rank-order data, analogous to principal components analysis for interval data (Bentler, 1970). Three-, four-, and five-factor solutions were obtained and rotated to an orthogonal, simple structure orthosim criterion (Bentler, 1977). Each item was then clustered into one of three groups on the basis of which factor it most adequately assessed. Since the three-factor solution proved to be most clearly interpretable, a target rotation was performed (Bentler, 1971) and the orthogonal loadings were interpreted (Bentler, 1968).

Results

Factor computations. Factor loadings for the three-dimensional solution are presented in Table 1; these loadings summarize the extent to which each item measures each dimension. Items have been grouped into three clusters on the basis of which factor they assess best.

Three clear-cut factors emerged, showing a strong correspondence to the hypothesized dimensions. Thus, parent perceptions of normal and hyperkinetic children vary in at least three ways, so that a particular child may show any combination of behaviors from the three dimensions; within a dimension, however, if a child tends to exhibit one of the behaviors, the child tends to exhibit the others. On the whole, items tend to measure a given factor quite well and are not relevant to other factors. (Contrast the low loadings

¹ Requests for copies of the rating scales should be sent to Arnold H. Zukow, 5363 Balboa Boulevard, Encino, California 91316.

Table 1
Parent Form: Dimension Loadings

Item	<i>M</i>	Factor loading		
		1	2	3
Factor 1: Excitability				
Irritable	.45	.76	.16	.16
Quick-tempered, explosive	.60	.73	.02	.05
Overly sensitive	.49	.71	.01	.30
Emotionally high strung	.44	.68	.11	.24
Unpredictable, unmanageable	.43	.63	.20	.27
Panics easily	.30	.56	.28	.21
Poor coordination	.24	.55	.15	.24
Can't seem to keep from touching everything and everyone around him	.45	.54	.29	.37
Fidgets	.63	.42	-.02	.36
Reacts adversely to changes in routine	.29	.41	.27	.07
Factor 2: Motor Coordination				
Has trouble drawing, writing	.19	.03	.80	.17
Has trouble buttoning	.21	-.19	.78	-.03
Eyes and hands don't seem to function together	.09	.47	.70	.14
Trouble with bicycle	.13	.08	.69	.24
Exceptionally clumsy	.10	.35	.65	.12

Item	<i>M</i>	Factor loading		
		1	2	3
Factor 2: Motor Coordination				
Was slow learning to walk	.16	-.02	.63	.19
Speech development has been slow	.15	.28	.50	.01
Factor 3: Directed Attention				
Is child lazy—not trying to do well in school?	.19	.18	.17	.82
Daydreams while doing homework assignments	.24	.20	.08	.82
Not learning in school although seems "bright"	.24	.14	.22	.77
Knows work orally at home—gets to school and has to write it down—fails.	.14	.16	.25	.75
Short attention span	.53	.31	.20	.59
Tolerance for failure and frustration is low	.64	.40	.08	.56
Jumps from one activity to another	.45	.24	.12	.52
Unusually hyperactive	.48	.45	.13	.49

Note. Those factor loadings that are considered to define a given factor appear in italics.

on irrelevant factors.) In some cases, items appear to be falling in between two factors, as indicated by a high loading on two factors; such items are ignored in factor interpretation.

The items defining the first factor appear similar to those marking hyperactivity as described by others. It is an excitability factor concerned with both the quality and quantity of behavior, including perpetual motion accompanied by explosive outbursts. The second factor refers to motor coordination exclusively. Items include general descriptions of poor coordination and more specific delays and difficulties with buttoning, walking, draw-

ing, speech, and so on. Finally, the third factor refers to directed attention or sustained participation in goal-directed behaviors. The items deal with motivation, attention/control, and learning as perceived by the parent. All item means are given in Table 1, which shows the relative proportion of children considered to exhibit a given behavior.

Factor scores. Once the three factors had been extracted, a score was computed for each child on each factor. The items whose loadings are italicized in Table 1 were used to compute three factor scores, with Factor 1 being measured by nine items, Factor 2 by seven items, and Factor 3 by seven items.

Scores were obtained by adding a point for every yes response. The three factor scores proved to be an adequately internally consistent, as measured by coefficient alpha, index representing the extent to which items are measuring the same dimension. Obtained alphas were .84, .74, and .82 for the three factors, respectively. To assess whether the three scores measured somewhat different aspects of behavior, the intercorrelations among the scales were obtained. In the combined sample of 160 children, the correlations were .30, .58, .30 among Factors 1 and 2, 1 and 3, and 2 and 3, respectively. These correlations are sufficiently low relative to the internal consistencies, so one can conclude that the three factors are reliably measuring distinct aspects of behavior. On the other hand, the significant and positive correlations among all three scores suggests that all items measure something in common, which, according to our experimental procedures, should involve the differentiation of hyperkinetic children from controls. Consequently, a fourth score was obtained as the sum of the previous three scores.

In addition, the two items of Table 1 not included in the three separate factor scores were included in this total score. Since "poor coordination" was deemed inconsistent with the interpretation of the other items marking Factor 1, that item was excluded from the computation of the factor total score. In view of the clear-cut nature of Coordination (Factor 2) and the absence of a high loading of the poor coordination item on Factor 2, it is apparent that parents interpreted this item as reflecting more globally on the quality of activity rather than to a specific motor task as might be expected, nor did parents interpret this item in terms of inattentiveness as reported by Conners (1969). The item "unusually hyperactive" loaded quite similarly on Factors 1 and 3 and therefore was not included in either factor score. Even though these two items do not clearly define a given factor, they do contribute to the overall description of hyperkinetic behavior. Thus these two items are included along with all italicized items of Table 1 in the computation of the fourth, total score. Its internal consistency

Table 2

Parent Form: Analyses of Variance

Factor	Source of variance			
	Subject status (A) ^a	Age (B) ^a	A×B ^a	Error ^b
1: Excitability				
<i>MS</i>	547.6	3.4	10.6	3.8
<i>F</i> ratio	143.0**	.9	2.8	
2: Motor Coordination				
<i>MS</i>	31.3	.0	1.6	2.2
<i>F</i> ratio	14.3**	.0	.7	
3: Directed Attention				
<i>MS</i>	250.4	10.5	26.8	2.2
<i>F</i> ratio	111.6**	4.7*	11.9**	
Total score				
<i>MS</i>	2,600.3	19.7	.2	11.3
<i>F</i> ratio	230.7**	1.8	.0	

^a *df* = 1.^b *df* = 156.* *p* < .05.** *p* < .001.

was .89, higher than the three separate scores, so that this score provides the most reliable overall description for these children. There are thus four scores for each child. The scores were examined for group mean differences.

Group comparisons. The parent form discriminates among clinical and control status at statistically significant levels. The statistical analyses are presented in Table 2. Each of the three factors, as well as the total score, verifies the statistically higher means obtained by clinical subjects as compared with control subjects. As expected, the total score yields the most reliable differentiation. In only one case was there a significant effect of age of subjects on the scores. This occurred in the third scale, concerned with directed attention. The effect is quite minimal in significance level compared to subject status, however, and can only be interpreted simultaneously with the interaction effect. Even though among the clinical group older subjects received higher scores, among the controls the age trend was minor and reversed.

Cross-validation. Results from preliminary work in cross-validation showed that among 26 clinical subjects (3½–14 years old),

all four factor scores on the parent rating scales were significantly different ($p < .001$) from the original control subjects. In all cases the direction of the differences and the size of the difference was virtually identical to that reported in the original study. A comparison of the original clinical group means to the means of the new clinical subjects revealed a high degree of similarity. More studies to substantiate this finding are indicated.

Classification. The task of differentiating clinical from control subjects can be accomplished by using the total scores, with cutoff scores being determined to maximize correct classification. Children receiving scores of 0–8 can be labeled as *normal*, whereas those scoring 9 or more can be labeled as hyperkinetic. Comparing the “true” status of the children, as determined by pediatrician judgments, with the labels obtained solely by the cutoff score yields the following picture: Ninety-six percent of the 90 young subjects were correctly classified, and 90% of the 70 older subjects were correctly classified. A more detailed look at the younger subjects reveals that 89% of the clinical subjects and 98% of the controls were correctly classified. Among the older subjects, 90% of the clinicals were correctly identified, as were 89% of the controls. These percentages are quite encouraging, compared to the 70%–80% figures reported in similar studies (Conners, 1970). Among the cross-validated clinical subjects, the cutting score of 8/9 correctly identified 81% of the children. This decrease in classification accuracy is a typical concomitant of cross-validation, but fortunately the results demonstrate that the instrument can be expected to be successfully applied in future studies.

Teacher Form

Method

Initial item pool. The rationale for the teacher form followed that of the parent form, but it was necessary to limit the size of the item pool due to time demands on teachers. A total of 15 items were devised to represent a broad range of school activities: tapping, motor skills, attention, and impulsive/explosive behavior. To introduce sufficient variance into the responses, teacher ratings were made on a 5-point scale.

Subjects. A new set of subjects, consisting of 78 controls and 36 clinical subjects, participated. Subject selection is described under *Parent Form*. An attempt was again made to sample a broad age spectrum. Sample sizes at ages 2½–5½ were 22 and 11 for controls and clinicals, respectively. At ages 5½–11, the corresponding sample sizes were 53 and 25.

Method of analysis. Steps analogous to those reported for the parent form were carried out for the teacher form. Since the data were adequately continuous rather than binary, a traditional least squares factor analysis was undertaken. Initial communality estimates were squared multiple correlations. Two different factor analyses were undertaken, a three-factor solution and a two-factor solution. The first three eigenvalues of the three-factor solution were 8.03, 1.15, and .82, suggesting that two factors or even one factor might be sufficient to account for the item intercorrelations. Rotation of the three-factor solution yielded two interpretable factors and a third unclear factor. The two-factor solution, rotated by orthosim, in contrast, yielded the same two interpretable factors.

Table 3
Teacher Form: Factor Loadings

Item	Factor loading	
	1	2
Factor 1: Attention/Excitability		
Is attention span short?	.83	.33
Child fidgets	.78	.30
Is the child a behavioral problem in class?	.78	.22
Unable to follow directions?	.73	.40
Quick-tempered, explosive	.67	.20
Seems to touch everything and everyone around him	.66	.31
Finds it hard to play with his peers	.65	.37
There are no activities that the child can focus his attention on	.62	.56
Has a low tolerance for failure and frustration	.61	.38
Reacts adversely to changes in routine	.60	.39
Factor 2: Motor Coordination		
Exceptionally clumsy	.20	.89
Coordination is poor	.32	.85
Speech development is slow or not clear	.16	.55
Eyes and hands can't seem to function together	.37	.53
Quiet and withdrawn—a loner	.19	.40

Note. Factor loadings that contribute to the total score for each factor appear in italics.

Results

Factor computations. Factor loadings for the two-factor solution are presented in Table 3. As before, items have been grouped into two clusters on the basis of which factor they best assess. The content of the items suggests that the first factor deals with attention and excitability, combining the two separate parent factors. It seems to represent the classical aspects of hyperactivity noted for the hyperkinetic syndrome. Factor 2, mirroring the content of the second parent factor, deals with motor coordination.

Factor scores. A score was determined for each child on each factor by summing the rating score given a child for each item in a given factor. A third total score was obtained analogously to the parent form by summing across the two factor scores and, in addition, including the item dealing with inability to focus continuously on any one activity; that item had been omitted from the two separate scores because it falls between the two factors. An internal consistency coefficient was obtained for each dimension, and the intercorrelation among the two factor scores was obtained in the combined total sample.

The internal consistency coefficients were again sufficiently high to reinforce the concept that the items comprising the two factor scores are measuring something in common. The alpha coefficients of .93 and .81 were compared to the intercorrelation of .63 between the scores on the two factors, which is again sufficiently low to consider the two factor scores to be measuring an independent aspect of behavior. On the other hand, the correlation of .63 is sufficiently high to assure that the scores measure the common tendency toward hyperkinesis, as indexed by coefficient alpha of .94 for the 15-item total score.

Group comparisons. The teacher form discriminates among clinical and control status at statistically significant levels. Factor scores are able to discriminate between clinical and control subjects. The subject status column of Table 4 indicates that the clinical means were significantly higher than

Table 4
Teacher Form: Analyses of Variance

Factor	Source of variance			
	Subject status (A) ^a	Age (B) ^a	A×B ^a	Error ^b
1: Attention/Excitability				
MS	1,675.1	110.9	195.3	44.7
F ratio	37.5**	2.5	4.4*	
2: Motor Coordination				
MS	130.1	.6	7.6	10.6
F ratio	12.3**	.1	.7	
Total score				
MS	3,172.1	138.6	311.2	90.3
F ratio	35.1**	1.5	3.4	

^a $df = 1$.

^b $df = 107$.

* $p < .05$.

** $p < .001$.

the control means in all of the comparisons at a high level of significance. There appears to be no age effect at all, and in only one case was the interaction of subject status and age statistically significant. As before, this interaction suggests that scores for clinical children increase with age, whereas those of control children do not.

Cross-validation. Results from the preliminary cross-validation of the teacher form paralleled the findings from the preliminary cross-validation of the parent form. Significant differences were found between the 26 new clinical subjects and the original control group ($p < .001$).

Classification. Children with a score of 35 or less can be classified as "normal"; those receiving a score of 36 or more can be classified as "hyperkinetic." Among the 33 younger subjects, 70% were correctly classified, whereas 85% of the 78 older subjects were appropriately grouped. The cutting score of 35/36 correctly classified 85% of the new clinical subjects.

Discussion

The factors that emerged in the parent and teacher forms indicated that both parents and teachers differentiated the excitability/attention versus motor coordination as-

pects of hyperkinesis. However, the analysis of the parent form revealed that parents differentiated excitability and directed attention as well. It is possible that this result is an artifact of the items used in the two studies. A careful analysis of the item pools indicates that the content concerned with daydreaming, not doing well in school, and so on, was omitted from the school form. In future versions of the instrument, item tapping this content will be included.

Although the factor loadings make it appear extremely likely that the factors from the two instruments are similar, there was no way in this study of verifying or disproving this hypothesis. Strict similarity can be verified by the demonstration of a high correlation between scores obtained on the parent form and the comparable scores on the teacher form. Unfortunately, the design of the current studies did not allow data to be obtained from both parent and teacher sources on the same subjects, so a correlation coefficient could not be computed. Even though a significant correlation between the parent and teacher factors might be expected, parents, teachers, and others actually have different sources of observational data about a child. It is quite possible that there may be little agreement regarding, say, concentration problems. The teacher may be in a better position to observe such characteristics in the classroom. A further study on this issue is clearly called for.

Langhorne et al. (1976) had the data to correlate composite teacher ratings, chart ratings, psychiatrist diagnosis, and parent ratings but, unfortunately, did not report on the extent of agreement that might be observed using such composite ratings. As pointed out before, since their analyses were based on an arbitrarily selected, very small subset of their data, it is not possible to rely on their conclusion that there is little or no agreement among sources of data. It is possible that among the many items discarded from their analysis, there may have existed numerous items showing high agreement across data sources.

The factors that did appear in this study were similar to those found by previous workers. The teacher form attention/excitability

factor was very similar to the hyperactivity factor reported by Blunden et al. (1974), as well as to the hyperactivity factor identified by Conners (1969). Results from the present parent form lend support to Conners' (1969) finding that an inattentiveness factor is distinct from impulsivity/hyperactivity. Further, our results show a clear separation between directed attention, including distractibility and concentration difficulties and poor motor coordination. Items from these two categories were not well represented in previous research, as indeed some were absent from the current teacher form. Obviously, items dealing with other content material, such as social behavior, also deserve further study and should be included in future research on this topic.

Numerous researchers have emphasized the disappointing results found in many studies investigating the effect of medication (Whalen & Henker, 1976) and/or psychological (Gittelman-Klein et al., 1976; O'Leary et al., 1976) or educational therapy (Douglas, 1974; Keogh & Margolis, 1976) on academic achievement. Some of these findings may be attributed to the problem of obtaining sufficient homogeneity among subjects, as recently reviewed in Langhorne et al. (1976). For example, the criteria for subject selection in studies of hyperkinetic children is often quite global (Keogh & Margolis, 1976; Rie et al., 1975). Furthermore, the variation of criteria for subject selection precludes generalization of results from study to study.

A preliminary investigation on the effectiveness of medication as a function of initial scores on the instruments yielded mixed and somewhat disappointing results. Of the 20 possible correlations, two were statistically significant. Statistically significant correlations of .68 and .75 ($p < .02$) were found between clinical rankings of medication effectiveness and Factor 1 of the teacher form and the total score in a sample of nine children under 5½ years of age. The children most responsive to medication were those initially considered by teachers to be most disturbed in excitability or in combination with poor coordination. Since there have been few reliable predictors of responsiveness to medication (Ross & Ross, 1976),

this encouraging result is being investigated further.

The total factor score derived from the parent and teacher rating scales permits the accurate identification of subjects, although it must be admitted that the issue of differential diagnosis also merits future research. Scores on the individual factors may provide a vehicle for investigating differential diagnosis and examining potential relationships between particular treatment conditions and changes in disapproved behavior, academic performance, cognitive abilities, and social skills. There is a striking similarity between the factors that emerged in this analysis and aspects of attentional disorders observed and described by several research groups. Douglas (1972, 1974) has characterized the facets of attention necessary for adequate academic performance as the ability to focus, to inhibit impulsive responding, and to sustain and organize attention; that is, to "stop, look, and listen"; Keogh and Margolis (1976) as coming to attention, problem solving, and maintaining attention; and Whalen and Henker (1976) as differences in ability to modify motor behavior to meet situational demands, in socially adaptive behaviors including impulsivity, and in sustained attention. Although the components of attention proposed by these researchers are not isomorphic and probably not entirely independent, the individual factor scores might perform the important function in future research of providing an objective basis for subject selection for various treatment conditions and for assessing changes in behavior and performance. The relationship between the parent and teacher rating scales and the Margolis-Keogh children's checking task is currently under investigation to provide further validation of the scales.

In conjunction with the researcher's or clinician's own observations, these instruments fulfill the function for which they were designed—to provide supplementary information and to serve as a guide for the conscientious clinician or researcher in the identification and treatment of hyperkinesis (Frankenburg, 1974).²

Reference Note

1. Grossman, H. J. *Current management of hyperactivity*. Paper presented at the meeting of the American Academy of Pediatrics, Washington, D.C., October 1975.

References

- Bentler, P. M. A new matrix for the assessment of factor contributions. *Multivariate Behavioral Research*, 1968, 3, 489-494.
- Bentler, P. M. A comparison of monotonicity analysis with factor analysis. *Educational and Psychological Measurement*, 1970, 30, 241-250.
- Bentler, P. M. Clustran, a program for oblique transformation. *Behavioral Science*, 1971, 16, 183-185.
- Bentler, P. M. Factor simplicity index and transformations. *Psychometrika*, 1977, 42, 277-295.
- Blunden, D., Spring, C., & Greenberg, L. M. Validation of the Classroom Behavior Inventory. *Journal of Consulting and Clinical Psychology*, 1974, 42, 84-88.
- Cantwell, D. P. Natural history and prognosis in the hyperactive child syndrome. In D. P. Cantwell (Ed.), *The hyperactive child: Diagnosis, management, current research*. New York: Spectrum Publications, 1975.
- Conners, C. K. A teacher rating scale for use in drug studies with children. *American Journal of Psychiatry*, 1969, 126, 152-156.
- Conners, C. K. Symptom patterns in hyperkinetic, neurotic, and normal children. *Child Development*, 1970, 41, 667-682.
- Conners, C. K. Symposium: Behavior modification by drugs: II. Psychological effects of stimulant drugs in children with minimal brain dysfunction. *Pediatrics*, 1972, 49, 702-708.
- Conners, C. K., Rothschild, G., Eisenberg, L., Schwartz, L., & Robinson, E. Dextroamphetamine sulfate in children with learning disorders. *Archives of General Psychiatry*, 1969, 21, 182-190.
- Conners, C. K., Taylor, E., Meo, G., Kurtz, M. A., & Fournier, M. Magnesium pemoline and dextroamphetamine: A controlled study in children with minimal brain dysfunction. *Psychopharmacology*, 1972, 26, 321-336.
- Daids, A. An objective instrument for assessing hyperkinesis in children. *Journal of Learning Disabilities*, 1971, 4, 499-501.
- Douglas, V. I. Stop, look and listen: The problem of sustained attention and impulse control in hyperactive and normal children. *Canadian Journal of Behavioural Science*, 1972, 4, 259-282.
- Douglas, V. I. Differences between normal and hyperkinetic children. In C. K. Conners (Ed.),

² A copy of the materials can be obtained from the first author.

- Clinical use of stimulant drugs in children.* New York: American Elsevier, 1974.
- Fish, B. Stimulant drug treatment in hyperactive children. In D. P. Cantwell (Ed.), *The hyperactive child: Diagnosis, management, current research.* New York: Spectrum Publications, 1975.
- Frankenburg, W. K. Selection of diseases and tests in pediatric screening. *Pediatrics*, 1974, 54, 612-615.
- Gittelman-Klein, R., et al. Relative efficacy of methylphenidate and behavior modification in hyperkinetic children. An interim report. *Journal of Abnormal Child Psychology*, 1976, 4, 361-379.
- Keogh, B. K., & Margolis, J. Learn to labor and to wait: Attentional problems of children with learning disorders. *Journal of Learning Disabilities*, 1976, 9, 276-286.
- Kupietz, S., Bialer, I., & Winsbert, B. G. A behavior rating scale for assessing improvement in behaviorally deviant children: A preliminary investigation. *American Journal of Psychiatry*, 1972, 128, 1432-1436.
- Langhorne, J. E., Jr., Loney, J., Paternite, E. C., & Bechtoldt, H. P. Childhood hyperkinesis: A return to the source. *Journal of Abnormal Psychology*, 1976, 85, 201-209.
- Laufer, M. W., & Denhoff, E. Hyperkinetic behavior syndrome in children. *Journal of Pediatrics*, 1957, 50, 463-473.
- Millichap, J. G. Drugs in the management of hyperkinetic children. *Journal of the American Medical Association*, 1968, 206, 1527.
- Myers, G. J., & Pless, I. B. Where's the hyperactive child going? *Pediatrics*, 1976, 57, 1.
- O'Leary, K. D., Pelham, W. E., Rosenbaum, A., & Price, G. H. Behavioral treatment of hyperkinetic children. *Clinical Pediatrics*, 1976, 15, 510-515.
- Rie, H. E. Hyperactivity in children. *American Journal of Diseases of Children*, 1975, 129, 783-789.
- Rie, H. E., Rie, E. D., Stewart, S., & Ambuel, J. P. Effects of methylphenidate on underachieving children. *Journal of Consulting and Clinical Psychology*, 1976, 44, 250-260.
- Ross, D., & Ross, S. *Hyperactivity: Research, theory, and action.* New York: Wiley, 1976.
- Sleator, E. K., & von Neumann, A. W. Methylphenidate in the treatment of hyperkinetic children. *Clinical Pediatrics*, 1974, 13, 19-24.
- Sroufe, L. A., & Stewart, M. A. Treating problem children with stimulant drugs. *National Educational Journal of Medicine*, 1973, 289, 407-413.
- Stewart, M. A., Pitts, F. N., Craig, A. G., & Dieruf, W. The hyperactive child syndrome. *American Journal of Orthopsychiatry*, 1966, 36, 861-867.
- Varni, J. W. *A cognitive behavior self-regulation approach to the treatment of the hyperactive child.* Unpublished doctoral dissertation, University of California, Los Angeles, 1976.
- Werry, J. S. Developmental hyperactivity. *Pediatric Clinicians of North America*, 1968, 15, 581-599.
- Werry, J. L., & Sprague, R. L. Hyperactivity. In G. Costello (Ed.), *Symptoms of psychopathology.* New York: Wiley, 1970.
- Whalen, C., & Henker, B. Psychostimulants and children: A review and analysis. *Psychological Bulletin*, 1976, 83, 1113-1130.

Received June 14, 1976 ■

Imagery Assessment Through Self-report: What Do Imagery Questionnaires Measure?

Merrill Hiscock
University of Texas, Austin

Assessment difficulties have impeded progress in evaluating the therapeutic role of visual imagery. Four studies examined imagery questionnaires and addressed the issues of (a) reliability; (b) agreement among different questionnaires; (c) social desirability; and (d) construct validity. The Betts and Gordon scales, and a newer inventory, the Paivio Individual Differences Questionnaire, were examined. Reliability of the Paivio inventory was found to be satisfactory and equivalent to that of other imagery questionnaires. In two studies, the Betts and Paivio questionnaires were correlated at the .45-.50 level, but correlations involving the Gordon scale were inconsistent from one study to the next. In general, imagery measures were not influenced by social desirability. Factor analysis indicated that subjective and objective measures of visualization are independent. The final study revealed a relationship between imagery questionnaire scores and reported values and interests. It is suggested that imagery is not a unitary construct and that criteria other than visuospatial tests may be appropriate for validating imagery questionnaires.

The emphasis on imagery in otherwise dissimilar therapeutic techniques (e.g., Horowitz, 1970; Wolpe & Lazarus, 1966) suggests that understanding imagery may be as useful in clinical practice as in the learning laboratory (cf. Paivio, 1971). Behavior therapy techniques, in particular, frequently include instructions directing the client to imagine specified aversive or reinforcing stimuli. Substituting imagery instructions for in vivo presentation of stimuli is more than a matter of convenience, for imagery instructions potentially enhance the flexibility and power of various techniques (Cautela, 1971). However, it is not known to what extent imagery actually medi-

ates the obtained results; research has failed to establish a relationship between therapy outcome and self-reported individual differences in imagery (Beere, 1972; Davis, McLemore, & London, 1970; McLemore, 1972). Either therapeutic success is independent of ability to form images or the expected relationship is obscured by inadequate measurement of individual differences in imagery.

The assessment problem is difficult. It is not clear what imagery questionnaires really measure or what criteria are appropriate for validating them. Objective tests of spatial ability appear to share little or no variance with visual imagery questionnaires (Di Vesta, Ingersoll, & Sunshine, 1971; Forisha, 1975; Neisser, 1970; Paivio, 1971, p. 496), and no consistent relationship has been established between questionnaires and objective measures of memory for objects (cf. Danaher & Thoresen, 1972; Marks, 1972; Rehm, 1973; Rimm & Bottrell, 1969). In fact, it has been argued that the subjective experience of imagery bears little relationship to any measurable capacity (Neisser, 1970). Reliable physiological correlates of subjectively reported imagery are equally elusive (see Paivio, 1971, 1973; Richardson, 1969).

The author is indebted to Regina Hilbertz, Steve Jackson, Randy Routon, and Charlotte Wilson for their assistance in data collection; to David B. Cohen and an anonymous reviewer for their helpful comments and suggestions; to Lyle Freedman for making available the questionnaire data from the group of 114 subjects in Study 1; and to Ginger Hiscock and Leota Servetnyk for their assistance in preparing the manuscript.

Requests for reprints should be sent to Merrill Hiscock, who is now at the Division of Psychology, University Hospital, Saskatoon, Saskatchewan, Canada S7N 0W8.

The questionnaires themselves have received relatively little critical attention despite their probable psychometric shortcomings. Rating scales such as Sheehan's (1967b) version of the Betts Questionnaire upon Mental Imagery (reprinted in Richardson, 1969) and Marks's (1973) Vividness of Visual Imagery Questionnaire require subjects to judge their images using a numerical scale on which one pole always represents "good" imagery. This format is very susceptible to response sets (see Mischel, 1968). The same criticism can be leveled at the Gordon (1949) scale of imagery control (reprinted in Richardson, 1969) on which a yes response always indicates controllable imagery. In fact, Di Vesta et al. (1971) found that the factor that loaded on the Betts and Gordon questionnaires also loaded on the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960). This finding is troublesome, since it suggests the possibility that any relationship between therapeutic outcome and self-reported imagery merely reflects an underlying response set that influences both imagery reports and response to therapy.

The present research addresses four fundamental issues regarding imagery questionnaires: (a) internal consistency and retest reliability; (b) agreement among different subjective measures; (c) the influence of social desirability; and (d) construct validity. Attention is focused on an inventory—Paivio's (1971) Individual Differences Questionnaire (IDQ)—as a possible alternative to the traditional rating scale.

Study 1

Method

Subjects. As in the three subsequent studies, subjects were undergraduate students enrolled in introductory psychology courses at the University of Texas at Austin. Three groups of subjects participated in Study 1. One group comprised 48 females, the second comprised 114 subjects with an approximately equal number of females and males, and the third group comprised 43 females and 36 males.

Procedure. The first steps taken were to evaluate and improve Paivio's IDQ, which is an 86-item true-false instrument requiring respondents to report their characteristic "ways of thinking, studying, and problem solving." The questionnaire yields two scores: one for imagery and one for verbal processes. The directionality of scaling for several items is reversed to minimize the effect of agreement or acquiescence.

The IDQ was completed by the first two groups of subjects. Item-scale correlations were computed to identify nonproductive items; 15 items whose correlation with the respective scale score was less than .25 for both groups was excluded from the revised questionnaire. After 1 new item was added, the revised questionnaire contained 38 verbal and 34 imagery items. The format was changed to a Likert scale format with five available responses per item.

The 72-item IDQ was administered to the third group of subjects, and Cronbach's alpha (Cronbach, 1951) was computed to assess internal consistency. Fifty-eight (73%) of the 79 subjects completed the questionnaire a second time 2–6 weeks after initial testing. The scores of these subjects provided a basis for estimating retest reliability.

Results

Scores for both IDQ scales were distributed in Gaussian manner. Cronbach's alpha for the Imagery scale of Paivio's original questionnaire was .80 for the first group and .81 for the second group of subjects. The same statistic for the 72-item version was .87 for the group of 79 subjects. The Verbal scale yielded similar values of alpha: .83 and .86 for the original questionnaire and .88 for the 72-item version. Retest reliability coefficients for the 72-item IDQ were .84 for the Imagery scale and .88 for the Verbal scale.

Study 2

Method

This study addressed two issues: (a) the degree of association among the Paivio IDQ, Sheehan's version of the Betts Questionnaire upon Mental Imagery, and the Gordon scale of (visual) imagery control and (b) the correlation between each of these instruments and the Marlowe-Crowne Social Desirability Scale. A sample of 123 undergraduates (68 females, 55 males) completed the 72-item IDQ, the Visual and Auditory scales of the Betts questionnaire, and the Gordon scale. All but 10 of these subjects also completed the Marlowe-Crowne Social Desirability Scale. Correlations were computed.

Results

Table 1 shows the correlations among the five imagery scales and between each of these scales and the Marlowe-Crowne scale. Sex of subject is also included in the intercorrelation matrix; negative correlations with sex reflect higher scores for females. The polarity of cor-

Table 1
Correlations Among Imagery Questionnaires and the Marlowe-Crowne Social Desirability Scale

Variable	2	3	4	5	6	7
1. Sex	.11	-.29**	-.04	-.33**	.03	-.20*
2. Gordon scale		.16	.10	.21*	.17	-.10
3. Betts Visual scale			.40**	.49**	.17	.09
4. Betts Auditory scale				.21*	-.15	.10
5. Paivio Imagery scale					.13	.02
6. Paivio Verbal scale						-.11
7. Marlowe-Crowne scale						

Note. $n = 123$ for all variables except the Marlowe-Crowne scale, for which $n = 113$.

* $p < .05$.

** $p < .01$.

relations involving the Betts questionnaire has been reversed to compensate for the negative scaling of that instrument. Positive associations between imagery reported on the Betts scales and other variables thus appear as positive correlations.

The strongest association indicated in Table 1 is between the Paivio IDQ Imagery scale and the Betts Visual scale ($r = .49$). Correlations between the Gordon scale and the other two visual imagery scales did not exceed .21. There was a moderately strong association between the Visual and Auditory scales of the Betts questionnaire ($r = .40$) but virtually no association between the Visual and Verbal scales of the Paivio IDQ. The Betts Auditory scale and the IDQ Verbal scale were not related.

No variable showed a correlation greater than .20 with the Marlowe-Crowne Social Desirability Scale. The variable that did correlate at the .20 level with social desirability was sex; females tended to score higher than males. Correlations between the Marlowe-Crowne scale and the five imagery scales did not exceed .11 in magnitude, and their average was exactly zero.

Correlations of $-.29$ and $-.33$ between sex and the Betts Visual scale and between sex and the Paivio IDQ Imagery scale, respectively, indicate that females report "better" imagery than do males. The means were 10.62 for females versus 12.87 for males on the Betts scale, $t(121) = 3.36$, $p < .01$, and 134.99 for females versus 126.02 for males on the IDQ, $t(121) = 3.77$, $p < .01$. Separate intercorrelation matrices were computed for females and males, but the only significant difference involved the

relationship between the Gordon scale and the Marlowe-Crowne scale. The correlation for females was $-.28$ versus .11 for males ($z = 2.05$, $p < .05$).

Study 3

Method

The 43 females and 36 males who completed the 72-item Paivio IDQ in Study 1 were given six other questionnaires and tests. Among these were the Gordon scale and the Visual and Auditory scales of the Betts questionnaire. An objective test of spatial ability, the Minnesota Paper Form Board Test, was included in the battery. This test, which has been used by Ernest and Paivio (1969, 1971) as part of an "imagery battery," represents a possible exception to the usual finding that objective tests are unrelated to imagery questionnaires (cf. Paivio, 1971, p. 496). The Quick Word Test (Borgatta & Corsini, 1960) was selected as a vocabulary test. Because of time constraints, only the first 35 items were used. The final two measures were developed specifically for use in the present study. One of these, called the Visual Memory Scale, requested subjects to imagine six common objects in the local area and then "read out" specific information from their images. For example, subjects were asked to visualize the university tower and count the columns of windows on one face. The other measure, called the Visual Manipulation Scale, required subjects to mentally manipulate various objects. A representative item from this scale, taken from Griffiths (1927), instructs subjects to picture a 3-inch (4.62 cm) cube, painted red, that is sawed into 1-inch (1.54 cm) cubes. Subjects were asked to determine the number of little cubes having paint on three faces.

All together, 10 variables were represented in the intercorrelation matrix. The three imagery questionnaires contributed a total of five scales; there were four objective tests, and subject's sex was the 10th variable. The intercorrelation matrix was factor analyzed, and the factors extracted were rotated using a varimax procedure.

Table 2
Correlations Among Subjective and Objective Measures

Variable	2	3	4	5	6	7	8	9	10
1. Sex	-.10	-.36**	-.02	-.19	-.16	.21	.10	.11	-.02
2. Gordon scale		.47**	.37**	.56**	.20	-.09	.06	-.05	-.06
3. Betts Visual scale			.43**	.46**	.05	-.16	.01	-.09	-.09
4. Betts Auditory scale				.24*	-.04	-.05	-.14	-.12	-.08
5. Paivio Imagery scale					.09	-.06	.04	-.14	-.03
6. Paivio Verbal scale						.03	.12	.30**	-.12
7. Visual Memory Scale							.02	.21	.03
8. Visual Manipulation Scale								.39**	.28*
9. Quick Word Test									.18
10. Minnesota Paper Form Board									

Note. $n = 79$.

* $p < .05$.

** $p < .01$.

Results

The intercorrelation matrix is presented in Table 2. Again, negative correlations with sex reflect higher scores for females, and positive correlations with the Betts scales reflect positive associations.

Relationships among the five imagery scales were similar to those found in Study 2, except that correlations between the Gordon scale and the two other visual scales were markedly stronger than those found earlier. The correlation between the Gordon scale and the Betts Visual scale ($r = .47$) was significantly greater than .16 in Study 2 ($z = 2.38$, $p < .05$), and the correlation between the Gordon scale and the Paivio IDQ Imagery scale (.56) was significantly greater than .21 in Study 2 ($z = 2.86$, $p < .01$). The IDQ Imagery scale and the Betts Visual scale again were correlated at a moderate level ($r = .46$).

Females again tended to report "better" imagery than males on the IDQ Imagery scale and the Betts Visual scale, but only the Betts scale yielded a significant sex difference, $t(77) = 3.30$, $p < .01$. The mean for females was 10.81 versus 14.28 for males. When separate intercorrelation matrices were calculated for females and males, only two significant differences were found, and both involved the Gordon scale. The correlation between the Gordon scale and the Betts Visual scale was significantly greater for males ($r = .66$) than for females ($r = .08$, $z = 2.99$, $p < .01$). Similarly,

the correlation between the Gordon scale and the Betts Auditory scale was greater for males ($r = .60$) than for females ($r = .00$; $z = 2.92$, $p < .01$).

There was no appreciable association between subjective and objective measures of imagery. The Paivio IDQ Verbal scale, however, did show a modest correlation with the Quick Word Test ($r = .30$).

Four factors with eigenvalues greater than 1.0 emerged from factor analysis. Loadings for each of these factors are shown in Table 3. Factor 1 loaded on the three visual imagery scales and the Auditory scale of the Betts questionnaire; Factor 2 loaded on the Minnesota Paper Form Board and the Visual Manipulation Scale; Factor 3 loaded on the subject's sex and the Visual Memory Scale; and Factor 4 loaded on the Paivio IDQ Verbal scale and the Quick Word Test.

Study 4

Method

This study further explored the pattern of correlations involving the Paivio IDQ. Eighty-one males completed the IDQ and three other tests. Space Relations, from the Differential Aptitude Tests (Bennett, Seashore, & Wesman, 1966), was used as an objective test of spatial ability. Like the paper form board, Space Relations is a component of the "imagery battery" used by Ernest and Paivio (1969, 1971). The Quick Word Test again served as a vocabulary test. This time all 100 items were administered.

Table 3
Factor Loadings for Four Varimax Factors

Variable	Factor				h^2
	I	II	III	IV	
Betts Visual scale	.75	-.01	-.31	.03	.66
Gordon scale	.82	.02	-.01	.16	.69
Paivio Imagery scale	.75	.04	-.13	.05	.58
Betts Auditory scale	.68	-.15	.15	-.22	.56
Minnesota Paper Form Board	-.07	.79	-.09	-.27	.71
Visual Manipulation Scale	.04	.76	.08	.25	.65
Sex of subject	-.14	.05	.79	-.20	.69
Visual Memory Scale	-.05	-.01	.70	.17	.52
Paivio Verbal Scale	.09	-.05	-.09	.88	.78
Quick Word Test	-.11	.50	.29	.56	.67
Eigenvalue	2.49	1.69	1.17	1.16	

The fourth test in the present study was the Study of Values (Allport, Vernon, & Lindzey, 1960). The imagery literature contains some evidence that characteristics of an individual's imagery are related to variables such as interests and choice of vocation (Chowdhury & Vernon, 1964; Roe, 1951). On this basis, one might expect to find a relationship between the Paivio IDQ Imagery scale and the Study of Values, which is a 45-item ipsative scale yielding a score for each of six categories of interests and values. The scale can be dichotomized into two groups of values: extraceptive (theoretical, economic, and political) versus intraceptive (aesthetic, social, and religious) (Dunn, Bliss, & Siipola, 1958). It was predicted on the basis of the Dunn et al. findings that the latter categories would be associated with visual imagery (see Paivio, 1971). Thus, rank scores for these categories were computed and summed, so that subjects who gave high ratings to these three values were assigned a high numerical score.

Results

The intercorrelation matrix is displayed in Table 4. There was no significant association between the Paivio IDQ Imagery scale and Space Relations ($r = .05$), but there was a correlation of .35 between the IDQ imagery score and the rank-order index derived from the Study of Values. As predicted, subjects who reported more frequent use of imagery tended to have high aesthetic, social, and religious scores. Other correlations involving the IDQ Imagery scale were trivial.

Two other coefficients in the matrix are notable. The correlation of .41 between the IDQ Verbal scale and the Quick Word Test replicates a finding of Study 3. Also, the Quick Word Test and Space Relations were associated at the .29 level.

Discussion

The formal psychometric properties of Paivio's IDQ that were evaluated were found to be satisfactory. Even though reverse scaling of several items should have counteracted response sets that might inflate a reliability estimate, retest reliability appears to be at least comparable to that of the shortened Betts questionnaire (Sheehan, 1967a) and Marks's (1973) Vividness of Visual Imagery Questionnaire.

The instability of correlations involving the Gordon scale requires explanation, especially since McLemore (1976) recently reported very similar findings. Also, the only differences between the intercorrelation matrices for females and for males involved the Gordon scale. An estimate of retest reliability is not available; but split-half reliability of the Gordon scale is reasonably good ($r = .77$ for Study 2 and .84 for Study 3). A more probable source of difficulty is the marked skewness of the frequency distribution. A concentration of scores at the high end of the scale allows a relatively small number of low scores to exert disproportionate influence on correlations with other measures. Consequently, correlations vary between samples. Since the modal score is the maximal score (12 yes responses), transformations are not of much help. Moreover, the departure from normality will affect the bivariate distribution on which the product-moment correlation is based, and it may bias the z statistic used to assess the significance of differences between cor-

Table 4
Correlations among Space Relations, Study of Values, Quick Word Test, and the Paivio Individual Differences Questionnaire

Variable	2	3	4	5
1. Space Relations	.02	.29*	.05	.00
2. Study of Values		.08	.35*	.09
3. Quick Word Test			-.04	.41*
4. Paivio Imagery scale				.11
5. Paivio Verbal scale				

Note. $n = 81$ males.

* $p < .01$.

relations. Although the Gordon scale may be of some use in selecting cases for experimental or clinical purposes, it should either be modified (cf. Lane, 1975) or avoided in correlational and factor-analytic studies.

A milder degree of skewness in the distribution of scores on the Betts scales does not seem to affect this test's stability. On the contrary, correlations between the Betts scales and other variables (except the Gordon scale) proved to be remarkably stable from Study 2 to Study 3. Although the probable contribution of method variance cannot be overlooked (Mischel, 1968), the correlation between the Betts Visual scale and the IDQ Imagery scale suggests the existence of some subjective dimension that can be tapped using instruments that differ in format and content.

However, Di Vesta et al. (1971) concluded that subjective measures of imagery lack construct validity. There was no substantial correlation between imagery questionnaires (i.e., Betts's and Gordon's) and spatial tests, and the factor that loaded on the imagery questionnaires also loaded on the Marlowe-Crowne scale. Present data failed to support the Di Vesta et al. finding regarding social desirability, but perhaps the disparate evidence can be reconciled. First, Di Vesta et al. may have created a "method factor" by adding imagery questionnaires and the Marlowe-Crowne scale to a battery of objective tests. Second, even if the Betts Visual and Auditory scales are not contaminated by social desirability, some of the other Betts scales might be (cf. McLemore, 1976). Third, the general instability of correlations involving the Gordon scale could ac-

count for apparent differences in the influence of social desirability.

The present data support the Di Vesta et al. (1971) conclusion that imagery questionnaires and spatial tests are not interchangeable. Studies 3 and 4 demonstrate the kind of dissociation between self-reports and performance that led Neisser (1970) to argue that the accuracy or usefulness of imagery is unrelated to the subjective experience of imagery. It is especially interesting that the Betts Visual scale and the Visual Memory Scale were unrelated. Even subjects who reported the clearest and most vivid images were no more able than subjects on the other extreme of the Betts distribution to read out accurate information about visual aspects of familiar objects (cf. Woodworth, 1921, p. 372).

If imagery questionnaires are not contaminated by response sets, and nevertheless are unrelated to visuospatial performance, perhaps it is because imagery is not a unitary construct. That which imagery questionnaires measure may be different from, but no less interesting than, that which visuospatial tests measure. However, questionnaire validity still needs to be demonstrated. Two possibilities present themselves as potential criteria for validating imagery questionnaires. The first is performance on certain verbal memory tasks. Paivio (1971) has suggested that "an acquired disposition of the individual to react to words (especially concrete words) with nonverbal images" (p. 509) can be measured by questionnaires. Consistent with Paivio's conception the 72-item IDQ proved to be useful in predicting recall of high-imagery adjectives from prose (Hiscock, 1976). Study 4 suggests a second possible correlate of self-reported imagery. The obtained correlation of .35 between the IDQ Imagery scale and the Study of Values is not strong, but the relationship is consistent with previous findings (Chowdhury & Vernon, 1964; Dunn et al., 1958; Roe, 1951), and it demonstrates the need for more attention to values, interests, and choice of occupation as potential correlates of subjective imagery.

Two additional implications of the present research merit consideration. The first concerns the orthogonality of the Imagery and Verbal scales of the Paivio IDQ. Apparently, being a "verbalizer" is not simply the reciprocal

of being a "visualizer." For some purposes, especially studies of the relationship between imagery and language, the IDQ may prove most useful when it is used to select subjects on the basis of high scores on one scale and low scores on the other. The second observation pertains to sex differences. Females tend to score higher than males on the Betts Visual scale and the Paivio IDQ Imagery scale, and, in at least two imagery studies (Ernest & Paivio, 1971; Marks, 1973), females have outperformed males. Ernest and Paivio (1971) proposed that "in some tasks, females 'use' imaginal processes to facilitate recall whereas males do not" (p. 71).

No imagery questionnaire is likely to be the best choice for all applications. Paivio's IDQ promises to be a useful instrument for investigating habitual styles of information processing and, in particular, for investigating the manner in which people internally represent words. Consequently, it should be useful for studying the relationship of imagery to attitudes, interests, career choices, and so on. The IDQ thus offers a means of validating the concept of visual imagery as a cognitive style. The therapist, however, may prefer to have a "state" measure rather than a "trait" measure of imagery (McLemore, 1976). In this case, and also when nonvisual (e.g., tactile) imagery is of interest, a rating scale may be the only appropriate device.

References

- Allport, G. W., Vernon, P. E., & Lindzey, G. *Study of Values*. Boston: Houghton Mifflin, 1960.
- Beere, D. B. Systematic desensitization and imagination: A test of London's cognitive integration of behavior therapies (Doctoral dissertation, Michigan State University, 1972). *Dissertation Abstracts International*, 1972, 32, 5431B-5432B. (University Microfilms No. 72-8632, 234)
- Bennett, G. K., Seashore, H. G., & Wesman, A. G. *Differential aptitude tests*. New York: Psychological Corporation, 1966.
- Borgatta, E. F., & Corsini, R. J. The Quick Word Test. *Journal of Educational Research*, 1960, 54, 15-19.
- Cautela, J. R. Covert conditioning. In A. Jacobs & L. B. Sachs (Eds.), *The psychology of private events*. New York: Academic Press, 1971.
- Chowdhury, K. R., & Vernon, P. E. An experimental study of imagery and its relation to abilities and interests. *British Journal of Psychology*, 1964, 55, 355-364.
- Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
- Crowne, D. P., & Marlowe, D. A. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 1960, 24, 349-354.
- Danaher, B. G., & Thoresen, C. E. Imagery assessment by self-report and behavioral measures. *Behaviour Research and Therapy*, 1972, 10, 131-138.
- Davis, D., McLemore, C. W., & London, P. The role of visual imagery in desensitization. *Behaviour Research and Therapy*, 1970, 8, 11-13.
- Di Vesta, F. J., Ingersoll, G., & Sunshine, P. A factor analysis of imagery tests. *Journal of Verbal Learning and Verbal Behavior*, 1971, 10, 471-479.
- Dunn, S., Bliss, J., & Siipola, E. Effects of impulsivity, introversion and individual values upon association under free conditions. *Journal of Personality*, 1958, 26, 61-76.
- Ernest, C. H., & Paivio, A. Imagery ability in paired-associate and incidental learning. *Psychonomic Science*, 1969, 15, 181-182.
- Ernest, C. H., & Paivio, A. Imagery and sex differences in incidental recall. *British Journal of Psychology*, 1971, 62, 67-72.
- Forisha, B. D. Mental imagery verbal processes: A developmental study. *Developmental Psychology*, 1975, 11, 259-267.
- Gordon, R. An investigation into some of the factors that favour the formation of stereotyped images. *British Journal of Psychology*, 1949, 39, 156-167.
- Griffitts, C. H. Individual differences in imagery. *Psychological Monographs*, 1927, 39(1, Whole No. 172).
- Hiscock, M. Effects of adjective imagery on recall from prose. *Journal of General Psychology*, 1976, 94, 295-299.
- Horowitz, M. J. *Image formation and cognition*. New York: Appleton, 1970.
- Lane, J. B. Imagination and personality: The multi-trait investigation of a new measure of imagery control (Doctoral dissertation, University of Minnesota, 1974). *Dissertation Abstracts International*, 1975, 35, 6099B. (University Microfilms No. 75-12, 105)
- Marks, D. F. Individual differences in the vividness of visual imagery and their effect on function. In P. W. Sheehan (Ed.), *The function and nature of imagery*. New York: Academic Press, 1972.
- Marks, D. F. Visual imagery differences in the recall of pictures. *British Journal of Psychology*, 1973, 64, 17-24.
- McLemore, C. W. Imagery in desensitization. *Behaviour Research and Therapy*, 1972, 10, 51-57.
- McLemore, C. W. Factorial validity of imagery measures. *Behaviour Research and Therapy*, 1976, 14, 399-408.
- Mischel, W. *Personality and assessment*. New York: Wiley, 1968.
- Neisser, U. Visual imagery as process and as experience. In J. S. Antrobus (Ed.), *Cognition and affect*. Boston: Little, Brown, 1970.
- Paivio, A. *Imagery and verbal processes*. New York: Holt, 1971.
- Paivio, A. Psychophysiological correlates of imagery. In F. J. McGuigan & R. A. Schoonover (Eds.), *The*

- psychophysiology of thinking*. New York: Academic Press, 1973.
- Rehm, L. P. Relationships among measures of visual imagery. *Behaviour Research and Therapy*, 1973, 11, 265-270.
- Richardson, A. *Mental imagery*. New York: Springer, 1969.
- Rimm, D. C., & Bottrell, J. Four measures of visual imagination. *Behaviour Research and Therapy*, 1969, 7, 63-69.
- Roe, A. A study of imagery in research scientists. *Journal of Personality*, 1951, 19, 459-470.
- Sheehan, P. W. Reliability of a short test of imagery. *Perceptual and Motor Skills*, 1967, 25, 744. (a)
- Sheehan, P. W. A shortened form of Betts' questionnaire upon mental imagery. *Journal of Clinical Psychology*, 1967, 23, 386-389. (b)
- Wolpe, J., & Lazarus, A. A. *Behavior therapy techniques*. New York: Pergamon Press, 1966.
- Woodworth, R. S. *Psychology: A study of mental life*. New York: Holt, 1921.

Received November 8, 1976 ■

INSTRUCTIONS TO AUTHORS

Style of manuscripts. Manuscripts must be prepared in the style described in the *Publication Manual of the American Psychological Association* (2nd ed.). Instruction on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors should submit manuscripts in triplicate and should keep a copy to guard against loss.

Abstracts. Manuscripts of regular articles must be accompanied by an abstract of 100-175 words. Manuscripts of brief reports must be accompanied by an abstract of 75-100 words. All abstracts must be typed on a separate sheet of paper.

Brief Reports. The *Journal of Consulting and Clinical Psychology* will accept Brief Reports of search studies in clinical psychology. The procedure is intended to permit the publication of soundly designed studies of specialized interest or limited importance which cannot now be accepted as regular articles because of lack of space. Several pages in each issue will be devoted to Brief Reports, published in the order of their receipt without respect to the dates of receipt of the regular articles. Most Brief Reports appear in the first or second issue to go to press following their final acceptance.

An author who wishes to submit a Brief Report:

1. Sends the Brief Report, limited to two printed pages and prepared according to the specifications given below.

2. Agrees not to submit the full report to another journal of general circulation.

Specifications. The Brief Report should give a clear, condensed summary of the procedure of the study and as full an account of the results as space permits.

To ensure that a brief report does not exceed two printed pages, follow these instructions for typing: (a) Set typewriter to a 48-space line. (b) Type text. (c) Count all lines except abstracts, title, and by-line, including acknowledgments. If you have exceeded 185 lines, shorten the material. (d) If your brief report barely exceeds 75 lines (one printed page), try to edit to 75 lines exactly. Reports slightly exceeding one page will be edited for length in order to use space economically.

In Brief Reports, headings, tables, and references are avoided or, if essential, must be counted in the 75 lines. This journal no longer requires an extended report, but, if one is available, each Brief Report must be accompanied by the following footnote:

Requests for reprints and for an extended report of this study should be sent to John Doe (give the author's full name and address).

The footnote should be typed on a separate sheet and *not* counted in the 75-line quota.

A Model of Psychopathology Based on an Integration of MMPI Actuarial Systems

Harvey A. Skinner
Addiction Research Foundation,
Toronto, Ontario, Canada

Douglas N. Jackson
University of Western Ontario,
London, Canada

Relationships among Minnesota Multiphasic Personality Inventory (MMPI) code types from the Gilberstadt and Duker and the Marks, Seeman, and Haller systems were evaluated using an empirical clustering procedure. Three superordinate types were identified: neurotic (*Hs*, *D*, *Hy*), psychotic (*Sc*, *Pt*), and sociopathic (*Pd*, *Ma*), which form a theoretical model of psychopathology. Patients can be readily classified according to this typology and a graphic display of the results constructed. The importance of differentiating profile elevation, scatter, and shape is discussed when matching a given MMPI profile to the various code types. The data from the MMPI do not support the practice of highly differentiated classification within the three superordinate types.

In recent years there has been considerable interest in the development of actuarial systems for test interpretation in clinical settings. The Minnesota Multiphasic Personality Inventory (MMPI) systems of Gilberstadt and Duker (1965) and of Marks, Seeman, and Haller (1974) are two well-known examples. Potential advantages of such systems include economy of description, use of previous information and experience accumulated for specific code types, and the facility for automating test scoring and interpretation to save the clinician valuable time. Thus, given the MMPI profile for a patient, the clinician can determine if this profile matches a specific code type. The implicit assumption is that information garnered for this code type applies, more or less, to this patient.

The Gilberstadt and Duker (1965) system has 19 prototypes that were identified using

the "classic case" approach. Marks et al. (1974) developed 16 adult code types by examining profile configurations in a sample of psychiatric patients. Although each system was developed using a different rationale, one immediate question is the degree of overlap between the two approaches. A second issue concerns the actual steps used in classifying a profile. For example, would it be profitable to examine the independent contribution of profile elevation, scatter, and shape (Cronbach & Gleser, 1953; Skinner, 1977) as a means of improving classification hit rates? Payne and Wiggins (1968) found that either system only classifies approximately 28% of a psychiatric hospital sample, whereas joint application resulted in an overall hit rate of 49%. This percentage will vary apparently as a function of the educational level of the patient population.

This article investigates empirically the degree to which MMPI profiles can be conceptualized in terms of three superordinate MMPI types (neurotic, psychotic, and sociopathic). A second aim is to evaluate the degree to which the Gilberstadt and Duker (1965) and the Marks et al. (1974) code type systems can be integrated in terms of such superordinate types. These three prototypes can be conceptualized as "ideal type constructs" (Hempel, 1965, chap. 7) that form a theoretical model of psychopathology, as assessed by the MMPI. In

This research was supported by Canada Council Grant S75-0495 IX to the University of Western Ontario and by the Addiction Research Foundation of Ontario.

The authors thank Lewis R. Goldberg and Richard I. Lanyon for providing punched data for MMPI groups.

Requests for reprints should be sent to Harvey A. Skinner, Addiction Research Foundation, 33 Russell Street, Toronto, Ontario, Canada M5S 2S1.

addition, this theoretical model offers certain practical advantages for classification research in clinical settings. The end result for classified patients is a set of elevation, scatter, and shape parameters that can be computed readily with a small calculator. Furthermore, graphic plots of the classification data can be constructed to facilitate interpretation.

Method

Subjects

The 35 MMPI code type profiles, 19 from Gilbertstadt and Duker (1965) and 16 (adult) from Marks et al. (1974), were keypunched for computer processing. Then, relationships among these 35 code types were examined using a multivariate classification strategy. Thus, the 35 MMPI code types formed the derivation sample for a new MMPI taxonomy.

To provide an estimate of the generalizability of this new MMPI taxonomy, data taken from Lanyon (1968) were classified. Lanyon has compiled mean MMPI profiles (*T* scores) for 233 diverse psychiatric and normal groups.

Procedure

Conceptually, one begins by hypothesizing that a set of ideal personality types (Hempel, 1965, chap. 7) underlie the two MMPI codebook systems. The basic concept is a modal profile, which can be defined as a hypothetical MMPI profile pattern that is characteristic of a subset of patients in a psychiatric population. Our

approach is based on the assumption that patients can be differentiated into relatively homogeneous classes by grouping those individuals who substantially resemble the same MMPI modal profile.

A least squares estimate of the modal profiles is derived through a generalized principal-components model. Jackson and Williams (1975) and Skinner (1977, in press; Skinner, Note 1) have provided discussions of this classification procedure from somewhat different perspectives. Briefly, consider a data matrix X_i giving the scores of N individuals from sample i on the 13 MMPI clinical scales. The classification model can be depicted as

$$X_i = f\{M_i, S_i, A_i P' + E_i\}. \quad (1)$$

That is, relationships among MMPI profiles (row vectors) of sample i can be expressed as a function of (a) a column vector of profile elevation parameters M_i , (b) a column vector of scatter parameters S_i , and (c) a matrix of shape parameters A_i describing the extent to which each individual's MMPI profile resembles the shape of the MMPI modal profiles. Note that shape describes the actual pattern of "ups and downs" across the 13 MMPI clinical scales and elevation depicts the degree to which the MMPI profile as a whole has high (or low) *T* scores, whereas scatter represents how dispersed the 13 scale scores are about their average (Cronbach & Gleser, 1953). Finally, E_i is a residual matrix that provides a measure of how well the modal profiles "fit" the particular sample X_i .

The present study focuses first on estimating the underlying "ideal types" or modal profiles defined by the matrix P . The 35 Code Types \times 13 MMPI Clinical Scales data matrix X was decomposed according to the Eckart and Young (1936) theorem (i.e., $X = F \Delta P'$). A least squares estimate of the modal profiles is provided by the right-hand eigenvectors P . Then, given the

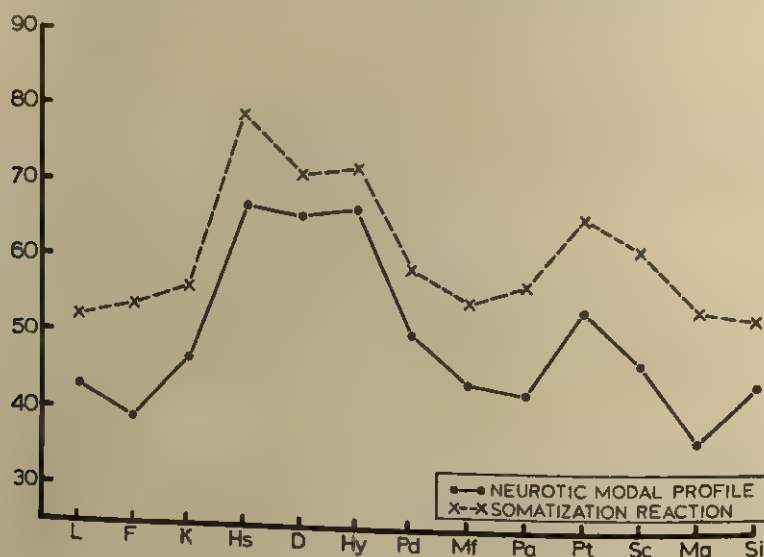


Figure 1. Neurotic modal profile and a somatization reaction group Minnesota Multiphasic Personality Inventory profile.

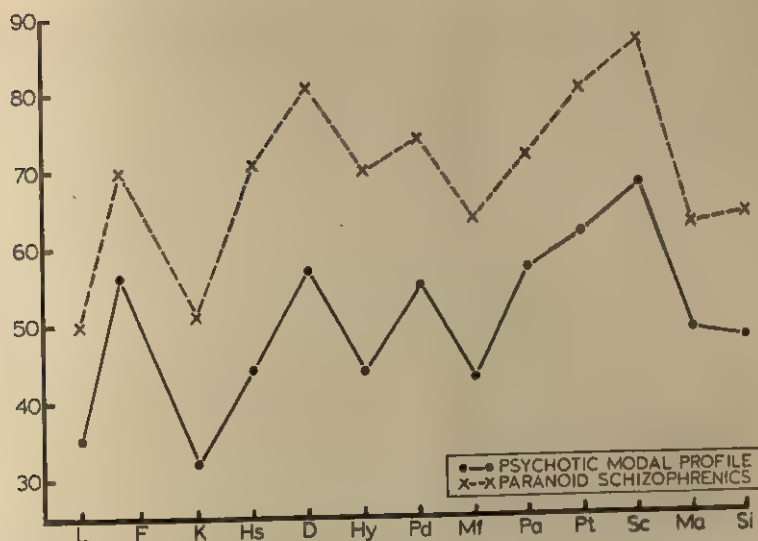


Figure 2. Psychotic modal profile and a paranoid schizophrenic group Minnesota Multiphasic Personality Inventory profile.

modal profiles *P*, the second aim of this study is to determine the degree to which the modal profiles are representative of the 233 Lanyon (1968) MMPI groups. Basically, one is fitting the model described in Equation 1 to the Lanyon data, in which each group MMPI profile is considered as a single entity.

In a previous study using this classification model, eight modal profiles based on a structured inventory of psychopathology were identified and replicated across three samples of alcoholic patients (Skinner, Jackson, & Hoffmann, 1974). The generalizability of these modal

profiles to diverse psychiatric and normal populations was explored by Skinner, Reed, and Jackson (1976).

Results

Essentially, the analysis considered similarity in profile shape among the 35 MMPI code types to identify a set of most representative MMPI profiles. Three modal profiles were derived (Table 1) and are plotted in Figures 1-3.

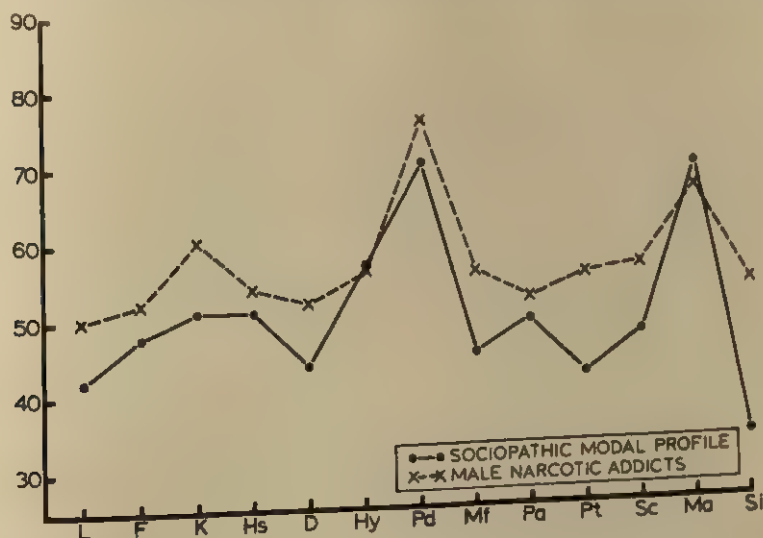


Figure 3. Sociopathic modal profile and a male narcotic addicts group Minnesota Multiphasic Personality Inventory profile.

Table 1
*Minnesota Multiphasic Personality
 Inventory Diagnostic Prototype
 Modal Profiles*

Scale	Modal profile		
	1	2	3
<i>L</i>	-.670	-1.509	-.820
<i>F</i>	-1.060	.596	-.171
<i>K</i>	-.256	-1.832	.093
<i>Hs</i>	1.716	-.580	.053
<i>D</i>	1.612	.655	-.563
<i>Hy</i>	1.665	-.561	.747
<i>Pd</i>	.059	.493	2.043
<i>Mf</i>	-.561	-.683	-.467
<i>Pa</i>	-.732	.705	-.068
<i>Pt</i>	.396	1.153	-.757
<i>Sc</i>	-.311	1.832	-.305
<i>Ma</i>	-1.255	-.125	1.886
<i>Si</i>	-.603	-.145	-1.671

Note. Each modal profile has a mean = .0 and standard deviation = 1.0. Table 1 corresponds to Matrix *P* in Equation 1.

In Table 1 the data have a mean of 0 and variance of 1, whereas each modal profile is scaled in the figures to have a mean of 50 and a variance of 10. Note that the modal profiles in Table 1 correspond to the matrix *P* of the general model depicted in Equation 1. Table 2 presents the various code types with their elevation, scatter, and three shape parameters. The first modal profile, with high points on *Hs*, *D*, and *Hy*, classifies the neurotic code types from each system. For example, the Gilberstadt and Duker (1965) codes 1-3-2 and 1-2-3-7, and the Marks et al. (1974) code 2-3-1 are the purest exemplars of this *neurotic* ideal type. The second modal profile has high points from *Sc* and *Pt*. The best code type exemplar for this *psychotic* ideal type is the code 8-6 from each system. Finally, the third modal profile, labeled *sociopathic*, is characterized by *Pd* and *Ma*. Not surprisingly, the code 4-9 from each system has the highest weight.

The weightings in Table 2 describe the correlation of each codebook type with the three modal profiles of Table 1. Thus, these weightings emphasize conformance in profile *shape*, that is, the actual pattern of ups and downs across MMPI scales (Matrix *A*, in Equation 1). A second important aspect of profile resemblance is *elevation*, which is defined as the pa-

tient's average across the 13 MMPI scales (Vector *M*, in Equation 1). Note that it is possible for two individuals to have a profile shape quite similar to a modal profile depicted in Figure 1, even though the two individual profiles may be widely different in elevation. For example, a psychiatric patient and a normal subject may show a high degree of similarity to the neurotic modal profile, indicating respective high points on *Hs*, *D*, and *Hy*. However, these two individuals may be differentiated at a second stage when considering elevation. One would expect the psychiatric patient to show more severe psychopathology and thus to have high points close to or exceeding a *T* score of 70.

Also depicted in Figures 1, 2, and 3 is the mean MMPI profile for a reference group taken from Lanyon (1968). These reference group profiles provide "real world" exemplars of the three ideal types. The somatization reaction profile plotted in Figure 1 represents a group of 39 white veterans from the psychiatry service of the Minnesota Veterans Administration Hospital (Rosen, 1958). Only "pure" cases were selected for this diagnostic category. The MMPI profile for this group has a marked similarity in shape ($r = .93$) to the neurotic modal profile, although the somatization reaction profile is higher in elevation. Similarly, Figure 2 presents the MMPI profile of 100 white male veterans from the psychiatry service of the Minnesota Veterans Administration Hospital who had been diagnosed as paranoid schizophrenics (Rosen, 1958). This profile substantially resembles the shape ($r = .92$) of the psychotic modal profile. However, the paranoid schizophrenics' profile is higher in elevation. Finally, Figure 3 presents a male narcotic addict's profile based on adult addicts treated at the Patton (California) State Hospital (Olson, 1964). This reference group profile closely resembles the shape ($r = .82$) of the sociopathic modal profile and is slightly higher in profile elevation.

Classification results for selected group profiles from Lanyon's (1968) compendium are given in Table 3.¹ The three modal profiles accounted for 68.79% of the variance due to pro-

¹ Data for the complete set of 233 groups are available from the authors.

Table 2
Classification Results

Classification Results			Shape		
System	Elevation	Scatter	1	2	3
Gilberstadt & Duker (1965)					
	62.61	13.22	.94	.00	-.01
Code 1-2-3	65.69	12.72	.90	.26	.27
Code 1-2-3-4	64.69	14.09	.96	.15	-.09
Code 1-2-3-7	59.31	12.09	.96	-.16	.07
Code 1-3-2	65.39	12.48	.74	.32	.07
Code 1-3-7	67.54	11.16	.73	.40	.10
Code 1-3-8	58.31	8.71	.54	.00	.66
Code 1-3-9	63.85	12.94	.80	.40	-.20
Code 2-7	66.92	12.26	.59	.66	.18
Code 2-7-4	74.00	16.37	.44	.82	-.26
Code 2-7-8	58.00	7.52	.16	.22	.79
Code 4	60.31	8.97	.61	.09	.63
Code 4-3	55.54	9.04	-.16	.06	.95
Code 4-9	70.23	13.36	.47	.83	-.07
Code 7-8	81.15	19.83	.67	.47	-.03
Code 8-1-2-3	68.15	12.50	.30	.86	.27
Code 8-2-4	70.61	14.26	-.08	.89	-.04
Code 8-6	63.31	11.28	-.34	.72	.42
Code 8-9	53.92	9.79	-.54	.21	.68
Code 9					
Marks, Seeman, & Haller (1974)					
	63.54	12.07	.91	.30	-.10
Code 2-3-1	66.15	12.53	.78	.40	-.29
Code 2-7	67.54	12.35	.41	.73	-.03
Code 2-7-4	70.31	14.40	.55	.71	-.34
Code 2-7-8	72.39	13.65	.43	.87	-.10
Code 2-8	63.23	10.12	.85	-.17	.27
Code 3-1	67.31	13.25	.86	.40	.07
Code 3-2-1	64.23	11.25	.08	.66	.40
Code 4-6	67.54	14.73	.43	.73	.09
Code 4-6-2	69.61	11.91	.19	.88	.30
Code 4-8-2	61.08	8.37	-.04	.34	.90
Code 4-9	68.31	10.14	.62	.56	.27
Code 8-3	72.31	16.39	.03	.93	.04
Code 8-6	65.61	11.76	-.17	.83	.30
Code 8-9	57.79	9.20	-.50	.46	.56
Code 9-6	54.31	3.85	.34	-.21	.60
Normal					

Note. The elevation parameters M_i are the mean of the 13 clinical scale T scores, scatter parameters S_i are the profile standard deviation, and shape parameters A_i are the correlation between a code type and respective modal profile of Table 1.

file shape among the 233 Lanyon (1968) groups. Furthermore, using an entrance criterion of a highest loading above .50, 84.55% of the 233 groups can be classified as salient on a specific modal profile (cf. Skinner et al., 1976). These findings provide encouraging support for the generalizability of the three MMPI modal profiles among diverse deviant and normal samples. Note that classification data analogous to those contained in Table 3 can be readily com-

puted (see computational summary below) for any sample of patients with scores on the 13 MMPI clinical scales.

Given the classification results exemplified by Table 3, graphic displays can be constructed to facilitate interpretation of the data. That is, the elevation (M_i) and shape (A_i) parameters provide coordinates on orthogonal axes (Skinner, 1977, in press; Skinner, Note 1). For example, Figure 4 plots elevation by shape for

Table 3
Classification Results for Selected Groups from Lanyon

Lanyon (1968) group	Elevation	Scatter	Shape		
			1	2	3
Neurotic modal profile	50.00	10.00	1.00	.00	.00
Somatization reaction (15b)	60.92	8.19	.93	.17	.10
Somatization reaction (14a)	58.51	6.79	.92	.02	.17
Multiple sclerosis (39)	62.05	9.54	.95	.09	.04
Neurotic outpatients (13)	61.74	9.45	.87	.36	.11
Neurotic inpatients (13)	59.43	7.75	.88	.40	.05
Psychotic modal profile	50.00	10.00	.00	1.00	.00
Paranoid schizophrenics (2)	69.08	10.46	.38	.92	.02
Schizophrenics 15-29 (6a)	62.59	7.10	.14	.92	.13
Schizophrenics 30-39 (6a)	58.36	3.51	-.10	.85	.29
Acute schizophrenics (1)	64.15	6.26	.27	.86	.19
Chronic schizophrenics (1)	65.39	8.36	-.08	.78	.06
Sociopathic modal profile	50.00	10.00	.00	.00	1.00
Male narcotic addicts (21)	56.82	6.80	-.11	.03	.82
Female narcotic addicts (21)	56.05	6.64	-.21	.53	.72
Psychopathic personality (23)	58.20	6.38	.00	.56	.72
Prisoners (65a)	59.15	6.26	.22	.47	.76
Model prisoners (65b)	59.15	5.39	.49	.46	.61

Note. The number in parentheses after the group name refers to the figure number in Lanyon's (1968) book.

the psychotic modal profile. The horizontal axis represents the degree to which each code type from Table 2 (or any new MMPI profile) matches the *shape* of Modal Profile II. The vertical axis depicts the degree of profile *elevation*. Along with selected code types, two schizo-

phrenic group profiles (Lanyon, 1968) are plotted in Figure 4. Although the two profiles substantially resemble the shape of Modal Profile II, the schizophrenic 15- to 29-year-old group was lower in profile elevation than the paranoid schizophrenic group. Figure 5, on the

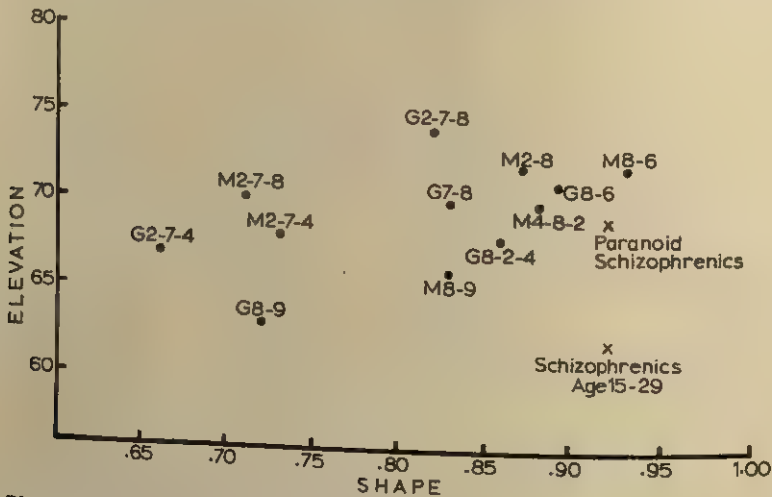


Figure 4. Plot of shape versus elevation for the psychotic modal profile. (G denotes a Gilberstadt and Duker, 1965, code type, and M designates a Marks, Seeman, and Haller, 1974, code type.)

other hand, displays the degree to which selected code type and Lanyon (1968) group profiles resemble the shape of Modal Profile I (neurotic) versus Modal Profile II (psychotic). Both figures provide a convenient summary of the classification analysis.

Discussion

This new method of identifying an MMPI taxonomy can be readily used in applied settings. That is, given a patient's MMPI profile in *T*-score format, the computational steps for forming Equation 1 are quite straightforward:

1. Compute the patient's average over the 13 MMPI clinical scales *T* scores to yield the elevation parameter (M_i).
2. Compute the standard deviation of the 13 MMPI scales about this patient's average to give the scatter parameter (S_i).
3. Compute the correlation of the patient's profile with each of the three modal profiles (Table 1) to yield the three shape parameters (A_i).

The investigator should scan the three shape parameters to identify the highest weighting. If this value exceeds a minimum standard (e.g., .50), then a figure for the appropriate modal profile (cf. Figure 4) should be constructed. The position of this patient can be located, given the elevation and shape coordinates. If, for example, a patient's location is most proximal to code 1-3-2 from Gilberstadt and Duker (1965), then the clinician could interpret the body of information regarding code 1-3-2 with respect to this patient. Furthermore, a secondary (tertiary, etc.) classification is possible by examining the location of this patient relative to other code types in the ordination space. Finally, one can interpret the theoretical and empirical data accumulated for the appropriate superordinate modal profile.

A technical issue concerns the degree of reliable differentiation among code types proposed by each system. Note that the three modal profiles of Table 1 explain (reproduce) 81.92% of the covariation among the 35 code types. That is, with three underlying "ideal type constructs," one is able to capture most of the common variance among the Gilberstadt and Duker (1965) and Marks et al. (1974) systems.

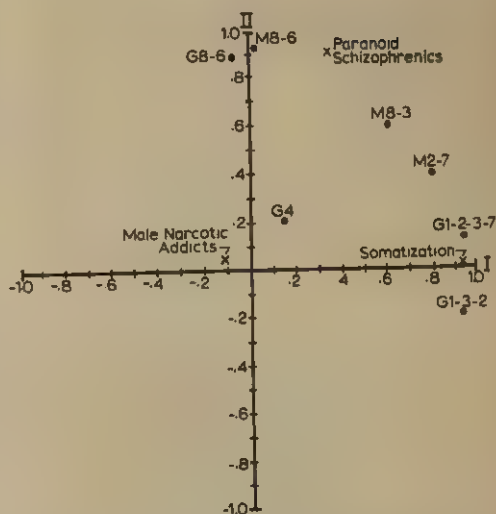


Figure 5. Plot of the neurotic (I) versus the psychotic (II) modal profile. (G denotes a Gilberstadt and Duker, 1965, code type, and M designates a Marks, Seeman, and Haller, 1974, code type.)

Certainly, one could question whether two code types within one system that are close together in the ordination space (e.g., Marks et al.'s, 1974, codes 2-8 and 8-6 of Figure 4) represent reliable differences. They should probably be collapsed into one code type.

From a theoretical perspective, it is hypothesized that the shape parameters yield indices of the patient's more enduring personality disposition (Lorr, 1966), whereas elevation and scatter reflect more temporary or situational factors influencing the degree of maladjustment (Carlson, 1972; Mori & Krane, 1973; Quertin, 1966). By examining the independent contribution of elevation, scatter, and shape components of profile similarity (Skinner, 1977, in press), one can integrate various aspects of the trait (core) and social-learning (situational) theory approaches to the study of personality (Bowers, 1973). For example, if pretreatment and posttreatment data are available, one could fit the MMPI modal profiles to each data set and then examine the temporal stability of the elevation, scatter, and shape parameters. One might hypothesize that patients will be classified in the same modal profile on each occasion when considering shape. However, a decrease in the elevation parameter could be predicted corresponding to

a reduction in symptom severity in response to a successful treatment program.

Although relatively homogeneous subgroups or clusters can be identified in the ordination space (cf. Figures 4 and 5), the current approach is essentially a dimensional model. The level of differentiation implicit to the three superordinate modal profiles is similar to Goldberg's (1972) hierarchical classification (i.e., normal vs. deviant; neurotic vs. psychotic vs. sociopathic).

In conclusion, the three MMPI modal profiles form a typological model of psychopathology that is open to empirical evaluation. Hopefully, this model will guide further theoretical developments and empirical research in the study of psychopathology. Expanding the system to more than three modal profiles will require additional psychopathological groups using more differentiated and less highly correlated personality scales than those comprising the MMPI clinical scales. Furthermore, by building on the actuarial systems of Gilberstadt and Duker (1965) and Marks et al. (1974), the present model should provide a convenient and useful diagnostic basis for clinical decision making.

Reference Note

1. Skinner, H. A. *Modal profile analysis and classification research*. Paper presented at the annual meeting of the Psychometric Society, Bell Laboratories, Murray Hill, New Jersey, April 1-3, 1976.

References

- Bowers, K. S. Situationism in psychology: An analysis and a critique. *Psychological Review*, 1973, 80, 307-336.
- Carlson, K. A. Classes of adult offenders. A multivariate approach. *Journal of Abnormal Psychology*, 1972, 79, 84-93.
- Cronbach, L. J., & Gleser, G. G. Assessing similarity between profiles. *Psychological Bulletin*, 1953, 50, 456-473.

- Eckart, C., & Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1936, 1, 211-218.
- Gilberstadt, H., & Duker, J. *A handbook for clinical and actuarial MMPI interpretation*. Philadelphia, Pa.: Saunders, 1965.
- Goldberg, L. R. Man versus mean: The exploitation of group profiles for the construction of diagnostic classification systems. *Journal of Abnormal Psychology*, 1972, 70, 121-131.
- Hempel, C. G. *Aspects of scientific explanation*. New York: Free Press, 1965.
- Jackson, D. N., & Williams, D. R. Occupational classification in terms of interest patterns. *Journal of Vocational Behavior*, 1975, 6, 269-280.
- Lanyon, R. I. *A handbook of MMPI group profiles*. Minneapolis: University of Minnesota Press, 1968.
- Lorr, M. *Explorations in typing psychotics*. Oxford, England: Pergamon Press, 1966.
- Marks, P. A. Seeman, W., & Haller, D. L. *The actuarial use of the MMPI with adolescents and adults*. Baltimore, Md.: Williams & Wilkins, 1974.
- Morf, M. E., & Krane, W. R. A strategy for developing automated assessment systems. *Behavioral Science*, 1973, 18, 294-306.
- Olson, R. W. MMPI sex differences in narcotic addicts. *Journal of General Psychology*, 1964, 71, 257-266.
- Payne, F. D., & Wiggins, J. S. The effects of rule relaxation and system combination on classification rates in two MMPI "cookbook" systems. *Journal of Consulting and Clinical Psychology*, 1968, 32, 734-736.
- Quertin, W. H. The search for recurring patterns among individual profiles. *Educational and Psychological Measurement*, 1966, 26, 151-165.
- Rosen, A. Differentiation of diagnostic groups by individual MMPI scales. *Journal of Consulting Psychology*, 1958, 22, 453-457.
- Skinner, H. A. "The eyes that fix you." A model for classification research. *Canadian Psychological Review*, 1977, 18, 142-151.
- Skinner, H. A. Differentiating the contribution of education, scatter and shape in profile similarity. *Educational and Psychological Measurement*, 1978, in press.
- Skinner, H. A., Jackson, D. N., & Hoffmann, H. Alcohol personality types: Identification and correlates. *Journal of Abnormal Psychology*, 1974, 83, 658-666.
- Skinner, H. A., Reed, P. L., & Jackson, D. N. Toward the objective diagnosis of psychopathology: Generalizability of modal personality profiles. *Journal of Consulting and Clinical Psychology*, 1976, 44, 111-117.

Received November 18, 1976 ■

Multimodal Evaluation of Practice Interactions as Treatment for Social Isolation

W. Stephen Royce and Hal Arkowitz
University of Oregon

Socially isolated college students, 28 men and 26 women, who volunteered for a program to increase social comfort and activity in friendship interactions, were randomly assigned to a treatment involving 12 real-life practice interactions with other subjects, a treatment involving 12 practice interactions plus 9 hours of social skills training, a minimal treatment control condition, or a delayed treatment control condition. Outcome was evaluated by multiple criteria that included self-report, self-monitoring, peer rating, and behavioral measures. Results indicated no significant differences between the two treatment groups or between the two control groups. The two treatment groups showed substantial and significant improvements in contrast to each control group on measures of social anxiety and social activity. These gains were maintained at follow-up assessments 3 and 15 months posttreatment. It is argued that the practice interaction treatment may function as *in vivo* desensitization, thereby reducing social anxiety and leading to increased social activity.

Recently, there has been a considerable amount of research on the assessment and treatment of minimal dating problems (Arkowitz, 1977; Curran, 1977). Surveys have indicated that social anxiety and minimal dating are significant and pervasive concerns for many individuals (e.g., Bryant & Trower, 1974). Several behavioral treatment procedures have been developed and evaluated for these problems, such as social skills training procedures involving behavior rehearsal, modeling, coaching, and feedback (e.g., Twentyman & McFall, 1975); anxiety-reduction procedures such as systematic desensitization (e.g., Curran & Gilbert, 1975); and cognitive modification

procedures (e.g., Glass, Gottman, & Shmurak, 1976).

Despite the considerable interest in minimal dating and the apparent efficacy of these different procedures, there has been very little attention directed toward another significant social problem—*anxiety and isolation in same-sex friendship interactions*. Friends provide opportunities for social comparison and information regarding social norms. In addition, they provide opportunities for modeling and feedback, which can have considerable influence in shaping a variety of social behaviors. Friendship relationships can also provide emotional support and comfort in times of stress. Thus, the lack of friendship relationships can severely limit an individual's social and emotional growth.

This article is based on a doctoral dissertation by the first author under the direction of the second author. The authors wish to thank Edward Lichtenstein, Susan Gilmore, Sheldon Cohen, and Dorothy Largay for their assistance in planning and executing this study. Gratitude is extended also to Kay Patterson, John Stanwood, Michael Shewin, and Jeff Lothman for their assistance in data collection and analysis.

Hal Arkowitz is now at the University of Arizona.

Requests for reprints should be sent to W. Stephen Royce, who is now at the Department of Psychology and Social Sciences, University of Portland, 5000 North Willamette Boulevard, Portland, Oregon 97203.

In the present study, a treatment approach that has been effective for minimal dating was applied to the problem of social anxiety and isolation in same-sex friendship interactions. The approach is based on real-life practice and involves repeated exposures to moderately anxiety-arousing social situations in the natural environment (Arkowitz, Christensen, & Royce, 1975). The procedure for minimal dating has involved providing subjects with a series of

six "practice dates," each with a different randomly selected partner who was also a volunteer for the program. Research has demonstrated that the procedure of practice dating leads to significant decreases in social anxiety and increases in posttreatment dating activity with new partners whom the subjects met on their own (Christensen & Arkowitz, 1974; Christensen, Arkowitz, & Anderson, 1975; Kramer, 1975). The effectiveness of the procedure appears to be due primarily to anxiety reduction (Christensen et al., 1975). An advantage of the practice dating procedure is that it can be administered on a large scale basis by someone trained only in clerical skills. In addition, since the treatment takes place in naturalistic situations, there is minimal concern about generalization from the therapy situation to the subjects' natural environments.

The procedure evaluated in the present study consisted of randomly pairing members of the same sex for a series of such "dates." This practice interaction group was compared to a second treatment group that received an identical series of practice interactions plus social skills training. (To the extent that subjects are characterized by inadequate social skills, the anxiety reduction effected through the practice interactions may not be sufficient for durable change to occur.) A minimal treatment group was used to control for attention and expectancy factors (Paul, 1969). A delayed treatment control group and a no-treatment group (to serve as a control for follow-up) completed the design. It was hypothesized that the two treatment groups would show significant improvements on multiple criteria of social skills, anxiety, and activity, relative to each of the control groups. It was also hypothesized that the practice interactions plus social skills training group would be superior to the group that received practice interactions only.

Method

Subjects

All subjects were college students, 18-23 years of age, who had telephones at their residences. The target population for the study was the most socially distressed 15% of approximately 1,000 students who completed a social survey questionnaire distributed in a number

of large classes. Since norms for social activity and anxiety in friendship interactions were not available, this selection criterion assured that subjects were considerably distressed and inactive.¹ Twenty-eight subjects were male, and 26 were female. Their mean age was 19.8 years, and 8% described themselves as married or cohabiting, 18% as engaged or steadily dating one person, and 74% as single.

Experimental Design

Subjects were randomly assigned among the two treatment groups and two control groups. Initial group sizes were 13 subjects in the practice-only group, 14 in the practice plus skill training group, 11 in the minimal treatment control group, and 16 in the delayed treatment control group. There was no mixing of sexes at any point in the study; males and females were treated separately but identically. Assessment was conducted pretreatment, posttreatment, 3 months posttreatment, and 15 months posttreatment. Because the delayed treatment control group was not available for control purposes at follow-up (they had received treatment by then), a follow-up control group was separately recruited.

Assessment

Self-report questionnaires. Two questionnaires were administered to all groups pretreatment, posttreatment, and at follow-up. The Social Avoidance and Distress Scale (Watson & Friend, 1969) is a 28-item social anxiety scale that has been demonstrated to possess good reliability and validity (Arkowitz, Lichtenstein, McGovern, & Hines, 1975; Watson & Friend, 1969). The self-report questionnaire consisted of free-response questions of retrospective frequency and range of same-sex social interactions and of dates and 7-point rating scales of same-sex social anxiety and social skills.

Self-monitoring measures. Subjects were asked to record their social interactions over a period of a week in a social interaction diary, which they carried with them. These diaries yielded a measure of the frequency and range (number of different people) of same-sex and opposite-sex interactions in a 1-week period. Interactions that did not progress beyond a mere greeting were not recorded. This procedure had been used by Christensen et al. (1975) in their practice dating study. The diary was administered pretreatment and posttreatment to all subjects except those in the follow-up control group.

Peer ratings. A peer questionnaire was sent to two male and two female acquaintances of each subject. The questionnaire yielded same-sex and opposite-sex social competence scores, based on ratings of social anxiety, social skills, and social activity. Because there

¹ For a more detailed description of the selection criteria used in this study, as well as a discussion of some general issues regarding selection criteria in studies of this type, see Royce and Arkowitz (1977).

seemed likely to be a substantial lag between changes in subjects' behavior and their peers' perceptions of these changes, this questionnaire was administered to all groups pretreatment and at the 3-month follow-up only.

Behavioral performance task. Subjects were brought into the laboratory where they engaged in a 10-min conversation with another subject. Four measures for each subject were taken from the audio recordings of the interactions: total talk time, number of silences longer than 10 sec, and ratings of social skill and social anxiety. These ratings were made by a corps of 22 untrained undergraduates. Five or six raters of the same sex as the subject independently rated each interaction. The means of these ratings yielded each subject's scores. This procedure, rather than the more usual one of training a small number of raters to a high degree of interrater agreement, was used to avoid the problem of relying solely on interrater agreement as the criterion of rating accuracy (Johnson & Bolstad, 1973; Lipinski & Nelson, 1974). The procedure yielded instead a representative peer opinion as to what constituted levels of social anxiety and skill. These measures were collected pretreatment and posttreatment from all subjects except those in the follow-up control group.

Process ratings. After each practice interaction, subjects completed ratings of themselves and their partners on 7-point scales of social anxiety and social skill.

Treatment

Practice only. Subjects were matched for 12 practice interactions, 2 per week, with a different same-sex partner each week. Matching was completely random. Each subject was sent the name and telephone number of his or her partner at the beginning of each week. Both interactions with that partner had to occur within a week, but decisions about all other details (initiation, time, place, length, etc.) were left up to the subjects.

Practice plus skill training. Subjects in this group participated in a series of 12 practice interactions as the practice-only subjects did. In addition, these subjects attended six weekly group social skills training sessions. Each group meeting lasted 1.5 hours. Except for the initial meeting of each group, the first 30 minutes of each session were devoted to feedback given to each subject by her or his partner from the previous week's interactions. The remaining hour was devoted to social skills training. Each week, subjects received a chapter from a social interaction training manual, based on similar manuals developed by McGovern (1972) and by Watkins (1972). Each chapter provided information about a particular topic in the realm of socially effective behavior and ended with a number of paper-and-pencil exercises to be completed by the subject during the week. Each group session was devoted to skills training in the area defined by the week's reading and exercises. Modeling and behavior rehearsal were the primary techniques used.

Minimal treatment control. Subjects in this group participated in six weekly group counseling sessions,

similar in format to the groups attended by the practice plus skill training subjects. These control subjects received the same readings and exercises as the practice plus skill training subjects, but group meetings consisted only of discussion and verbal counseling. These subjects received no practice interactions.

Delayed treatment control. Subjects in this group were telephoned and informed that the limitations of the program staff and facilities and the large number of volunteers for the program necessitated a delay in treatment for them. They were told that they would be able to participate in the program during the following academic quarter.

Follow-up control. Since the delayed treatment control subjects received treatment prior to any follow-up assessment, a separate group of subjects was recruited to serve as a no treatment control for follow-up. This group consisted of 13 male and 12 female students who had scored among the most distressed 15% on the screening questionnaire passed out in classes and who met the other screening criteria but who had not been contacted about participating in the treatment study. Data from these subjects were obtained only on the Social Avoidance and Distress Scale, the self-report questionnaire, and the peer questionnaire.

Procedure

Prospective subjects came in for a screening interview and to receive written and verbal explanations of the program. Interested and eligible subjects then signed a contract and consent form and paid a \$2 fee to cover expenses plus a refundable \$10 security deposit. During the third week of the winter academic quarter, pretreatment assessment was done. Treatment was given through the remainder of the winter quarter. Posttreatment assessment was administered during the second week of spring quarter, and subjects' \$10 deposits were then refunded. Delayed treatment control subjects were treated with practice plus skill training during spring quarter. Three months after treatment all subjects except those in the delayed treatment control group were paid \$1 to complete the follow-up questionnaires. The peer questionnaires were mailed again at that time. Fifteen months after treatment, practice-only and practice plus skill training subjects were again contacted and were asked to complete the follow-up questionnaires.

Results

All subjects completed the program, including posttreatment assessment, except for one practice-only subject, one practice plus skill training subject, and three delayed treatment control subjects. At the 3-month follow-up, there was an additional loss of two practice-only subjects, two practice plus skill training subjects, one minimal treatment control subject, and three follow-up control subjects. These losses were due to subjects'

lack of interest in further participation in the program.

One-way analyses of variance of pretreatment scores showed no significant differences among groups. In addition, the follow-up control group was compared to all other groups combined, and no significant differences were found.

Analysis of all data (except the process ratings and the 15-month follow-up data) was by analysis of covariance, with the pretreatment scores for each measure serving as the single covariate for each analysis. Since two-way (Sex of Subjects \times Treatment) analyses showed no significant sex or Sex \times Treatment interaction effects, sexes were combined and

one-way analyses were used. Consistent with the hypotheses stated earlier, analyses were by means of specific planned contrasts rather than by omnibus tests across all groups. Contrasts of practice only versus practice plus skill training showed no significant differences on any of the measures. Thus, there was no support for the hypothesis that social skills training would enhance the effects of the practice interactions procedure. Contrasts of minimal treatment control versus delayed treatment control versus follow-up control showed no consistent significant differences—Only 1 of the 24 contrasts reached the .05 level of significance. Detailed results are therefore reported only for the planned contrasts

Table 1

Means on Selected Outcome Measures for All Groups Pretreatment, Posttreatment, and at 3-Month Follow-up

Measure	Treatment group				
	P+	PO	MTC	DTC	FC
Social Avoidance and Distress Scale					
Pretreatment	13.7	14.8	13.1	13.6	10.0
Posttreatment	8.6	8.6	10.8	11.3	9.1
3-month follow-up	7.9	8.3	10.9		9.2
Social interaction diary					
No. same-sex interactions					
Pretreatment	24.8	20.7	41.8	33.5	
Posttreatment	42.2	39.8	42.1	34.4	
Range of same-sex interactions					
Pretreatment	13.8	14.1	16.5	19.7	
Posttreatment	20.4	23.5	16.5	14.6	
Peer questionnaire same-sex ratings					
Pretreatment	4.2	4.5	4.4	4.4	4.7
3-month follow-up	4.7	5.0	4.6		4.5
Self-report questionnaire					
Frequency with friends, last month					
Pretreatment	8.3	8.5	19.9	12.2	12.2
Posttreatment	9.3	14.7	12.1	10.1	15.4
3-month follow-up	13.5	9.4	12.6		16.4
Range with friends, last month					
Pretreatment	6.9	6.6	12.5	9.4	13.5
Posttreatment	6.0	9.0	12.1	9.6	17.1
3-month follow-up	9.2	8.7	10.1		13.2
Self-rated same-sex social anxiety					
Pretreatment	3.9	4.1	3.5	3.4	3.1
Posttreatment	2.4	3.3	3.5	3.4	3.4
3-month follow-up	2.1	2.9	3.9		3.2
Self-rated same-sex social skill					
Pretreatment	4.1	3.5	4.0	4.1	4.6
Posttreatment	4.9	4.5	4.0	4.5	4.5
3-month follow-up	4.9	4.3	4.1		4.2

Note. P+ = practice plus skill training; PO = practice only; MTC = minimal treatment control; DTC = delayed treatment control; FC = follow-up control.

of the practice-only and the practice plus skill training groups combined versus each control group.

Outcome Measures

Table 1 presents the means on selected outcome measures for all groups pretreatment, posttreatment, and at the 3-month follow-up.

Social Avoidance and Distress Scale. At posttreatment, the combined practice groups had improved significantly on this measure in contrast to each of the three control groups: $t(35) = -2.01$, $p < .05$, for the contrast with the minimal treatment control group; $t(37) = -2.26$, $p < .05$, for the delayed treatment control group; and $t(49) = -3.12$, $p < .01$, for the follow-up control group. At the 3-month follow-up, the practice groups continued to improve relative to the minimal treatment control group, $t(30) = -1.98$, $p < .05$, and the follow-up control group, $t(42) = -2.49$, $p < .01$. Thus, the treatment groups showed clear superiority on this measure over the control groups both immediately posttreatment and 3 months posttreatment.

Self-report questionnaire. Responses to the questions about social interaction frequency and range, and number of dates, showed no significant improvements for the combined practice groups when compared to the control groups. On ratings of social anxiety, the combined practice groups improved significantly in contrast to each of the three control groups: $t(34) = -2.50$, $p < .01$, for the contrast with the minimal treatment control group; $t(33) = -2.19$, $p < .05$, for the delayed treatment control group; and $t(48) = -3.16$, $p < .01$, for the follow-up control group. At the 3-month follow-up, these contrasts were maintained for the minimal treatment control group, $t(31) = -3.83$, $p < .01$, and for the follow-up control group, $t(41) = -3.46$, $p < .01$. On ratings of social skill, the combined practice groups improved significantly only in contrast to the minimal treatment control group, $t(34) = 2.29$, $p < .05$, and the follow-up control group, $t(48) = 2.35$, $p < .05$. Only the latter contrast was maintained at the 3-month follow-up, $t(41) = 2.59$, $p < .01$. There was, then, substantial and consistent improvement

on self-rated social anxiety for the treatment groups and minimal improvement on self-rated social skill.

Social interaction diary. For both frequency and range of same-sex social interactions, the combined practice groups improved significantly in comparison to the minimal treatment control group, $t(34) = 2.64$, $p < .01$, for frequency, and $t(34) = 1.77$, $p < .05$, for range, and the delayed treatment control group, $t(33) = 2.49$, $p < .01$, for frequency, and $t(33) = 2.64$, $p < .01$ for range. There were no significant contrasts for opposite-sex interactions. It should be noted that for each of the practice groups, less than 2% of subjects' posttreatment interactions were with other subjects. Thus, these substantial increases in social activity were not due merely to subjects' interactions with people they met in the program.

Peer questionnaire. On the same-sex peer rating scale, the combined practice groups improved significantly in contrast to the follow-up control group, $t(45) = 3.56$, $p < .01$. The contrast with the minimal treatment control group was in the predicted direction and approached statistical significance, $t(28) = 1.34$, $p < .10$. Analysis of the individual items on this scale showed that these results were due primarily to changes in ratings of social activity, somewhat to changes in ratings of social anxiety, and not at all to ratings of social skill. There were no significant contrasts for the opposite-sex scale.

Behavioral performance task. None of the four measures taken from the audio recordings of these interactions showed significant contrasts.

Process Ratings

Matched sample t tests were used to compare practice subjects' process ratings from the first 3 weeks of treatment with ratings made during the last 3 weeks. Subjects' ratings of their own anxiety showed a significant decrease over time, $t(17) = 2.33$, $p < .02$. Subjects' ratings of their own skill and of their partners' anxiety and skill also changed in the predicted directions but did not reach statistical significance.

15-month Follow-up

Questionnaires were returned by 19 of the 27 subjects in the two practice interaction groups. Matched-sample *t* tests computed on the pretreatment to 15-month follow-up data revealed that improvements noted at post-treatment and at the 3-month follow-up had been maintained. The mean score on the Social Avoidance and Distress Scale decreased to 7.2, $t(18) = -3.95$, $p < .01$; self-reported range of social interactions with friends increased to a mean of 13.6, $t(18) = 2.34$, $p < .02$; and self-rated same-sex social anxiety decreased to 3.0, $t(18) = -2.55$, $p < .01$. There were no significant effects for the other self-report items.

Discussion

The results indicate that the practice interaction procedure was an effective treatment for socially isolated college students. On multiple outcome measures of social anxiety and social activity, the practice interaction groups improved significantly in contrast to the control groups, and these gains were maintained at 3-month and 15-month follow-ups. The practice interaction procedure worked equally well for men and women. Data from the social interaction diary indicate that the vast majority of posttreatment social interactions were between subjects and persons whom they had not met through the program. This suggests that substantial learning occurred as a result of the program, and that this learning generalized to social interactions with other persons.

Treatment did not have any significant effects on any of the behavioral or rating measures of social skill, with the exception of the self-report questionnaire rating. The lack of change on measures of social skill for the practice-only group is consistent with findings from practice dating studies, which have also found no changes in social skill (e.g., Christensen et al., 1975). A surprising finding was that the addition of social skills training did not lead to any improvements on measures of social skill. Studies with minimal daters have found significant increases in social skills ratings as a result of skill training procedures

very similar to the ones used in the present study (Curran, 1977). Although unlikely, it is possible that the social skills training procedure was not a good one. Or, it may be that the measures of social skills were not sensitive to changes that did occur. However, a third possibility is that subjects were not deficient in social skills to begin with. If this were the case, then one would not expect any marked improvements in subjects' social skills, even with a viable social skills training program. It may be that the majority of subjects were characterized by excessive anxiety but adequate social skills. Several studies with minimal daters have supported this view (e.g., Clark & Arkowitz, 1975; Glasgow & Arkowitz, 1975). It may be that the same is true for individuals who are socially isolated with respect to friendship interactions. If this hypothesis were correct, it would also account for the finding that social skills training did not enhance the effects of the practice interaction procedure.

A most interesting question at this point concerns the mechanism by which the practice interaction procedure works. It may be that by "forcing" subjects to be socially active, they are exposed to the natural reinforcing properties of social activity, which maintains an increased rate of social behavior. In a similar manner, the subjects' social activity in the program may lead them to change their self-perceptions and labels of being shy and socially awkward. However, the explanation that seems most tenable at this point is that the practice interaction procedure functions as a means of *in vivo* desensitization, in which repeated exposure to moderately anxiety-arousing situations serves to reduce social anxiety and avoidance. Results of the present study give some support to this hypothesis: The process ratings demonstrated that self-rated social anxiety decreased from the first 3 weeks of practice interactions to the second 3 weeks, whereas skill ratings did not change. Christensen et al. (1975) found similar results. There is now a large and growing body of literature that demonstrates the effectiveness of *in vivo* exposure as a means of anxiety reduction. Marks (1975) reviewed over 300 studies of behavioral approaches to anxiety reduction and concluded that "real-life exposure is the

most powerful therapeutic factor so far identified" (p. 93).

Other questions also remain concerning this procedure as treatment for social isolation. Although the procedure has been found to be an effective treatment for college students' social inhibitions, these results may not generalize to other populations. The subjects in this study were student volunteers, not real clients. Even though these subjects were screened stringently to approximate a clinical population and to minimize the analogue nature of the study, a true clinical population may differ in important ways from the sample in this study. The procedure may not work with more severely distressed clients or with those less intelligent and less verbal than college students. Such individuals may also need specific social skills training.

References

- Arkowitz, H. The measurement and modification of minimal dating behavior. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 5). New York: Academic Press, 1977.
- Arkowitz, H., Christensen, A., & Royce, W. S. *Treatment strategies for social inhibition based on real-life practice*. Paper presented at the meeting of the Association for Advancement of Behavior Therapy, San Francisco, December 1975. (ERIC Document Reproduction Service No. Ed 120 605).
- Arkowitz, H., Lichtenstein, E., McGovern, K., & Hines, P. The behavioral assessment of social competence in males. *Behavior Therapy*, 1975, 6, 3-13.
- Bryant, B. M., & Trower, P. E. Social difficulty in a student sample. *British Journal of Educational Psychology*, 1974, 44, 13-21.
- Christensen, A., & Arkowitz, H. Preliminary report on practice dating and feedback as treatment for college dating problems. *Journal of Counseling Psychology*, 1974, 21, 92-95.
- Christensen, A., Arkowitz, H., & Anderson, J. Practice dating as treatment for college dating inhibitions. *Behaviour Research and Therapy*, 1975, 13, 321-331.
- Clark, J., & Arkowitz, H. Social anxiety and self-evaluation of interpersonal performance. *Psychological Reports*, 1975, 36, 211-221.
- Curran, J. P. Skills training as an approach to the treatment of heterosexual-social anxiety: A review. *Psychological Bulletin*, 1977, 84, 140-157.
- Curran, J. P., & Gilbert, F. S. A test of the relative effectiveness of a systematic desensitization program and an interpersonal skills training program with date anxious subjects. *Behavior Therapy*, 1975, 6, 510-521.
- Glasgow, R. E., & Arkowitz, H. The behavioral assessment of male and female social competence in dyadic heterosexual interactions. *Behavior Therapy*, 1975, 6, 488-498.
- Glass, C. R., Gottman, J. M., & Shmurak, S. H. Response acquisition and cognitive self-statement modification approaches to dating skills training. *Journal of Counseling Psychology*, 1976, 23, 520-526.
- Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts and practice*. Champaign, Ill.: Research Press, 1973.
- Kramer, S. R. Effectiveness of behavior rehearsal and practice dating to increase heterosexual social interaction. (Doctoral dissertation, University of Texas, 1975). *Dissertation Abstracts International*, 1975, 36, 913B-914B. (University Microfilms No. 75-16, 693).
- Lipinski, D., & Nelson, R. Problems in the use of naturalistic observation as a means of behavioral assessment. *Behavior Therapy*, 1974, 5, 341-351.
- Marks, I. Behavioral treatment of phobic and obsessive-compulsive disorders: A critical appraisal. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 1). New York: Academic Press, 1975.
- McGovern, K. B. Development and evaluation of a social skills training program for undergraduate male nondaters. (Doctoral dissertation, University of Oregon, 1972). *Dissertation Abstracts International*, 1973, 33, 4519B-4520B. (University Microfilms No. 73-7929).
- Paul, G. L. Behavior modification research: Design and tactics. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- Royce, W. S., & Arkowitz, H. Clarification of some issues concerning subject recruitment procedures in therapy analog studies. *Behavior Therapy*, 1977, 8, 64-69.
- Twentyman, C. T., & McFall, R. M. Behavioral training of social skills in shy males. *Journal of Consulting and Clinical Psychology*, 1975, 43, 384-395.
- Watkins, B. R. The development and evaluation of a transductive learning technique for the treatment of social incompetence. (Doctoral dissertation, University of Oregon, 1972). *Dissertation Abstracts International*, 1972, 33, 2361B. (University Microfilms No. 72-28, 199).
- Watson, D., & Friend, R. Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, 1969, 33, 448-457.

Received November 26, 1976 ■

Treatment Outcome and Follow-up Evaluation Based on Client Case Records in a Mental Health Center

Lynn S. Simons

Granite Community Mental Health Center,
Salt Lake City, Utah

Teru L. Morton

University of Hawaii

Terry C. Wade

Training and Career Development, State of Hawaii
Division of Mental Health, Honolulu

Dennis M. McSharry

Granite Community Mental Health Center,
Salt Lake City, Utah

The present study assessed the outcome of treatment of 121 mental health center clients using therapist and independent global improvement ratings and independent ratings of notes in case records based on a client-specific goal-oriented outcome technique (Goal Attainment Scaling; GAS). Telephone follow-up of 50 clients provided a second GAS assessment, client global improvement ratings, and three consumer satisfaction ratings. The findings indicated that (a) independently determined GAS scores and therapist and independent global ratings converged significantly, (b) the GAS procedure provided some increase in accuracy as well as an increase in specificity of outcome, and (c) client global ratings may reflect satisfaction with treatment rather than outcome. In view of the intercorrelations among measures and the relationship between GAS scores determined from case records and telephone interviews, case records may provide for accurate assessment of client problems and treatment success.

The measurement of treatment effectiveness in the applied setting is subject to many constraints (Twain, 1975). Clinicians often resist the "invasion" of accountability into their professional activities because they fear that their individual efforts are not effective, resent the time spent in evaluation that could be used for therapy, or believe that the clinical process is not quantifiable (Ellsworth, 1975). A level of evaluation beyond simple program monitoring may also require shifting resources from clinical activities to research, a source of both clinical and administrative resistance. Furthermore, considerations such as objectivity about the program being evaluated, level of competency of the evaluators, and potential usefulness of the evaluation results often argue in favor of external rather than internal evaluation (Weiss, 1972). External evaluation,

however, may increase costs and further increase resistance.

Methodological considerations also make treatment outcome research difficult (Bergin, 1971). Though Bergin and Suinn (1975) and Malan (1973) have concluded that the evidence for effectiveness of psychotherapy is relatively strong, there is great concern with the adequacy of the measurement of that effectiveness (Fiske et al., 1970; Garfield, Prager, & Bergin, 1971; Luborsky, Chandler, Auerbach, Cohen, & Bachrach, 1971; Strupp & Bloxom, 1975). Indeed, the primary focus of discussion in psychotherapy research has been the criterion of outcome measurement. The most frequently used measure has been the therapist rating of client global improvement (Luborsky et al., 1971), which seems to be the only criterion that shows consistent correlations with other measures of outcome (Cartwright, Kirtner, & Fiske, 1963; Fiske, Cartwright, & Kirtner, 1964; Strupp & Bloxom, 1975). However, this criterion tends to reflect a more positive view of outcome than specific change measures (Garfield et al., 1971), and it is certainly not

Requests for reprints should be sent to Lynn S. Simons, Research and Evaluation, Granite Community Mental Health Center, 156 Westminster Avenue, Salt Lake City, Utah 84115.

oriented toward differentiating areas in which change occurred (Cartwright, 1975). Other measures are subject to similar concerns for validity and utility and have prompted suggestions that multiple measures be used (Waskow & Parloff, 1975).

The aforementioned constraints on evaluation in the applied setting and concerns for treatment outcome measurement influenced the design of the present study. Based on suggestions by Fiske et al. (1970) and Luborsky et al. (1971) that criterion measures should be oriented toward the type of change or goals each patient needs or desires, the present study incorporated the client-specific, goal-oriented technique developed by Kiresuk and Sherman (1968). This technique (Goal Attainment Scaling; GAS) has been extensively used in various mental health, drug, criminal justice, rehabilitation, and education settings (cf. Garwick & Brintnall, 1974). In the present application, goals were established and attainment was measured retrospectively by independent raters who reviewed case records of former clients. This retrospective approach avoided potential resistance by eliminating therapist involvement in the evaluation, provided independent external evaluation, and represented a test of the utility of an innovative approach to studying treatment effectiveness at relatively low cost. Global improvement ratings by therapists and raters provided an indication of the correspondence between these more typical criteria and this novel outcome measurement approach (GAS).

In addition to outcome assessment at termination, a telephone follow-up of a sample of clients provided a second GAS score and client global improvement ratings to explore their utility as convergent outcome criteria. Clients were also asked three questions to see if their satisfaction with services might reflect improvement during treatment. Altogether, the present study used eight outcome measures in three content areas: GAS, global improvement, and consumer satisfaction.

Method

Setting

The present investigation was conducted at Granite Community Mental Health Center. The therapeutic orientation of the center is eclectic with a primarily

dynamic focus, and all traditional therapist categories are represented.

Observations

Terminated client case records ($N = 179$) were drawn in a 10% random sample from about 1,700 files that met two criteria: (a) Treatment was received from January 1, 1970, to June 30, 1975; and (b) the clients had completed at least 10 outpatient sessions, 8 partial hospitalization days, or 5 inpatient days. The second criterion was intended to insure that case records contained enough information to permit assessment of problems and outcomes using GAS. Despite this precaution, 58 of the case records were not amenable to outcome assessment for at least one of three reasons: inadequate intake notes (53%), inadequate progress notes (72%), and/or inadequate termination notes (91%).

Of the 121 measurable cases, 84 received outpatient treatment (mean of 26.6 sessions), 29 received partial treatment (24.5 days), and 8 received inpatient treatment (26.4 days). Fifty-three percent of the clients received primarily individual therapy, and 47% received primarily group or family therapy. (Only the record of the sampled family member was evaluated.) Four clients were treated by psychiatrists, 25 by psychologists, 59 by social workers or nurses, and 33 by paraprofessional therapists. Forty-eight percent of the clients were male. The average was 33.4 years (range = 11–70 years), and average formal education was 11.7 years. Thirty-four percent of the sample were single, 42% were married, and 24% were separated or divorced. Average family income was \$7,200 yearly, and average client family size was 3.96. Fifty percent of the clients had received some prior mental health treatment. The 188 problem areas identified for the 121 cases with measurable outcome were classified into 10 general categories: parent-child interactions (23.9%), depression (16.5%), marital problems (16.0%), other relationship problems (10.1%), self-understanding (9.0%), school problems (6.9%), anxiety (4.2%), agency referrals (3.7%), work problems (3.2%), and other problems (6.5%).

A comparison between clients for whom outcome could be assessed ($n = 121$) and for whom outcome could not be assessed ($n = 58$) revealed no significant differences in duration or type of therapy (individual or group or family), category of therapist, or client demographic characteristics. Fifty of the 121 clients whose records were measurable were reached for telephone follow-up. Of the 71 who could not be interviewed, 33.8% were not reached by a cutoff of three calls (which included the use of directory assistance to pursue changed numbers), 50.7% had moved or changed numbers that could not be traced, and 15.5% had no telephone listing. Again, there were no apparent treatment or client differences between those interviewed and those not interviewed.

Procedure

Two advanced graduate students in clinical psychology who had no affiliation with the Granite Com-

munity Mental Health Center contracted to conduct the retrospective outcome study and telephone follow-up. Both raters were trained in assessment and had some experience with GAS through designing evaluation programs and training staff in two applications of that technique.

Goal Attainment Scaling is a client-specific assessment and outcome technique developed by Kiresuk and Sherman (1968) to measure success in meeting treatment goals. It is used to (a) select one or more client problems during initial assessment, (b) establish possible outcome levels for each problem, and (c) compare actual outcome with established levels. Selected problems are each translated into verbal descriptions of possible outcome levels based on a common 5-point scale, which ranges from the least favorable treatment outcome thought possible (scale value = -2) to the most favorable outcome thought possible (+2). The expected outcome of treatment (scale value = 0) is based on clinical judgment about the most likely outcome for the individual client's particular problem. A goal attainment score with an expected value of 50 and an expected standard deviation of 10 can be computed for each client by combining scale outcome levels using a formula developed by Kiresuk and Sherman. A score of 50 indicates average expected treatment outcome, and a score of more than 50 indicates more than expected outcome. Computation of a single score for each client allows for comparisons of outcome between clients or programs.

As originally designed, the GAS technique requires therapist-client interviews for scale construction and outcome assessment, which can take up to 2 hours of client and therapist time. The retrospective nature of the present approach allowed scale construction and outcome measurement using only the closed case record. Scales were constructed from intake notes, and outcome was assessed from progress and termination notes. GAS scores were computed for each client's status at intake and termination, thus allowing measurement of changes during treatment. All problems in the intake notes that could be made into goals were scaled. Since their relative importance could not readily be determined, all goals were considered equally important for the purpose of computing GAS scores (cf. Kiresuk & Sherman, 1968).

Additional outcome measures included global ratings made by therapists, raters, and clients using 6-point scales (where 1 = considerably worse, 3 = unchanged, and 6 = considerably improved). Therapist global ratings were routinely made at termination and were available in the case records. Rater global ratings were made at the time termination GAS scales were assessed. Client global ratings were requested during telephone follow-up.

Telephone follow-up was conducted at an average of 2 years after treatment termination. The raters who interviewed the clients were unaware of clients' status at follow-up. Raters asked clients to indicate their status on each GAS by first reading them the termination level. (Clients were *not* told that this was the status recorded by their therapist at termination.) If clients described themselves as better or worse, the raters read the next scale level in the indicated direction.

They continued until the client said the appropriate level was reached, or the end of the scale was reached, whichever came first. Those levels for each client were then used to compute the follow-up GAS score.

Clients were next asked to rate the extent to which they felt they had improved since they first came to the Granite Community Mental Health Center for treatment using the global scale. Three additional questions were asked of each client to assess relationships between satisfaction with services and treatment outcome. Clients were asked to rate whether they would return to the center if they needed help in the future and whether they would recommend the center to others needing help on a 5-point scale (where 1 = definitely no, 3 = maybe, and 5 = definitely yes). Actual recommendation to others was scored dichotomously as the third measure.

Reliability

With any task involving rating, reliability of the ratings is an important issue. Prior to conducting the present study, the two raters spent approximately 8 hours each in training to apply the GAS technique to the case records. They determined subjective criteria (based on the amount and quality of information in the record) for deciding whether outcome was measurable, discussed goals and scale points for several cases, and independently developed and then discussed goal attainment scales for 12 records. Reliability assessment was conducted independently by the two raters on 27 cases of the 179 reviewed (13 initially reviewed by one rater and 14 by the other). There was complete agreement that 7 of the case records were not amenable to scale construction. For the 20 case records allowing outcome assessment, correlations of .94 and .93 were obtained between raters on termination GAS scores and global improvement ratings, respectively. These 20 case records allowed identification of 36 problem areas, for which there was 86% agreement between raters on the scale headings used to describe problems. The average termination outcome level for 121 clients was 48.86, with a standard deviation of 9.77. The similarity between the findings of the present study and the expected values of 50 and 10 suggest that the raters were correctly using the GAS technique.

Cost

One intent of the present study was to provide a large amount of outcome data at relatively low cost. The retrospective outcome portion of the study cost \$1,017 in direct rater time and indirect staff support at a cost per case reviewed of \$5.68. Excluding training and support time, the average time spent reviewing each of the 179 cases was 21 minutes. This compares very favorably with past efforts at outcome assessment at the Granite Community Mental Health Center. The telephone follow-up cost \$6.72 per case for the 50 clients who could be reached. The center administration has concluded that the benefits of the present study fully justify the cost in money and time that might otherwise have been devoted to clinical efforts.

Table 1
Correlation Matrix of Outcome Measures

Measure	1	2	3	4	5	6	7	8
1. Termination GAS score	9.77							
2. Rater global improvement	.80***	1.05						
3. Therapist global improvement	.52***	.45***	.90					
4. Follow-up GAS score	.40**	.36**	.18	13.96				
5. Client global improvement	-.08	-.03	-.14	.49***	1.12			
6. Would client return?	.12	.15	-.02	.27	.51***	1.12		
7. Would client recommend others?	.08	.15	-.06	.26	.50***	.74***	1.02	
8. Did client recommend others?	.00	.00	.08	.11	.50***	.28*	.27	.46

Note. Values in the diagonal are standard deviations. The measures are based on degrees of freedom of 119 for 1 and 2, 116 for 3, and 48 for 4-8. Significance tests are two-tailed. All correlations involving Measure 8 are point biserial. GAS = Goal Attainment Scaling.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Results

Both intercorrelations among outcome measures and the correspondence in magnitude of similar measures (e.g., global improvement ratings) obtained from different sources (e.g., therapists, raters, and clients) were of importance to the present study. The intercorrelations among measures are presented in Table 1. Notably, convergence among the major outcome measures was significant. Rater-determined termination GAS scores, $r(116) = .52$, $p < .001$, and rater global improvement ratings, $r(116) = .45$, $p < .001$, were both significantly correlated with therapist global improvement ratings. Likewise, termination GAS scores, $r(48) = .40$, $p < .011$, and rater global ratings, $r(48) = .36$, $p < .01$, were significantly correlated with client follow-up GAS scores. Moreover, only 1 of the 50 clients contacted at follow-up reported that the problem areas discussed were not those that brought them to treatment. However, the relationship between therapist and global ratings and follow-up GAS scores was not significant. Thus, GAS scores derived independently from the case record appear to provide a somewhat better index of client functioning than therapist global ratings, especially when both termination and follow-up are considered. It is also of interest to note that termination GAS scores and rater global ratings were highly correlated, $r(119) = .80$,

$p < .001$, which suggests that raters used similar criteria with these measures.

Client global ratings correlated significantly with follow-up GAS scores, $r(48) = .49$, $p < .001$, but not with therapist or rater global ratings or with termination GAS scores. Yet, client global ratings were substantially correlated with all three consumer satisfaction ratings (i.e., whether the client would return for treatment, whether the client would recommend others, and whether the client did recommend others). In fact, measures obtained during follow-up tended to be internally consistent but (with the exception of follow-up GAS scores) were independent of other measures. That is, client global ratings and the three consumer satisfaction ratings tended to correlate significantly with each other but not with other outcome measures.

With respect to the magnitude of measurements, the change in average GAS scores from intake (38.84) to termination (48.86) was highly significant, correlated $t(119) = 11.02$, $p < .001$. This finding confirms the sensitivity of the GAS procedure for revealing changes and its accuracy in approximating the expected level of 50. In addition, the 50 clients in the follow-up sample had a mean termination score of 49.54 and a follow-up score of 51.85, which were not significantly different. Clients seemed to improve during therapy and to maintain that improvement after therapy.

The magnitude of global ratings provided an additional perspective for evaluating the GAS procedure. Although the mean ratings of raters (4.06) and therapists (4.39) both indicated moderate improvement at the end of treatment (i.e., ratings of 3 = "unchanged" and 6 = "considerably improved"), therapists' ratings were significantly higher, correlated $t(116) = 3.84, p < .005$. Since there was high agreement between rater global ratings and termination GAS scores, the difference in mean ratings between raters and therapists suggests that therapist global ratings overestimate improvement (cf. Garfield et al., 1971).

Discussion

The present study evaluated the application of GAS to client case records as a measure of treatment effectiveness and examined its correspondence to other measures of outcome. The most important findings were that GAS scores derived from case records by independent raters at the termination of treatment converged significantly with therapist ratings of global improvement and GAS scores obtained from client reports at follow-up. Also, a comparison of GAS levels at intake and termination, together with the scores derived at termination and follow-up, indicated that the GAS procedure is sensitive to changes produced by treatment and is accurate in predicting expected outcome. In fact, results suggested that the GAS procedure is more accurate in measuring outcome (cf. Garfield et al., 1971; Luborsky et al., 1971). An important question, however, is whether raters can assess client problems from the case record. The significant correlation between GAS scores determined from the case record and GAS scores determined from a client interview seems to mitigate this concern. Additionally, only 1 client of the 50 interviewed reported that the problem areas being discussed were not those that brought the client in for treatment. Thus, it seems that independent raters can accurately assess problems and outcome from the case record.

Although client global improvement ratings appear to have face validity (i.e., they should reflect the client's experience with treatment),

and sometimes correlate with other outcome measures (Cartwright, 1975), client global ratings in the present study were unrelated to rater or therapist global ratings or GAS scores determined from case records. On the other hand, client global ratings were strongly related to whether the client would return for treatment, would recommend the center to others, and had recommended the center to others. Perhaps, when questioned about improvements during treatment, client reports largely reflect satisfaction with services rather than gains made.

Finally, it can be noted that the GAS procedure used in this study circumvented many problems frequently encountered in therapy outcome evaluation (cf. Ellsworth, 1975; Twain, 1975; Weiss, 1972), and provided a large amount of data at relatively low cost (cf. Fiske et al., 1970). In addition, dissemination of the generally positive results of this study have reduced staff resistance to evaluation and prompted center plans for further investigation, thereby carrying the intent of evaluation to utilization within the program evaluated (cf. Davis & Salasin, 1975).

References

- Bergin, A. E. The evaluation of therapeutic outcomes. In A. E. Bergin & S. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis*. New York: Wiley, 1971.
- Bergin, A. E., & Suinn, R. M. Individual psychotherapy and behavior therapy. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual Review of Psychology* (Vol. 26). Palo Alto, Calif.: Annual Reviews, 1975.
- Cartwright, D. S. Patient self-report measures. In I. E. Waskow & M. B. Parloff (Eds.), *Psychotherapy change measures*. Rockville, Md.: National Institute of Mental Health, 1975.
- Cartwright, D. S., Kirtner, W. W., & Fiske, D. W. Method factors in changes associated with psychotherapy. *Journal of Abnormal and Social Psychology*, 1963, 66, 164-175.
- Davis, H. R., & Salasin, S. E. The utilization of evaluation. In E. L. Streuning & M. Guttentag (Eds.), *Handbook of evaluation research* (Vol. 2). Beverly Hills, Calif.: Sage Publications, 1975.
- Ellsworth, R. B. Consumer feedback in measuring the effectiveness of mental health programs. In E. L. Streuning & M. Guttentag (Eds.), *Handbook of evaluation research* (Vol. 2). Beverly Hills, Calif.: Sage Publications, 1975.
- Fiske, D. W., Cartwright, D. S., & Kirtner, W. L. Are psychotherapeutic changes predictable? *Journal of Abnormal and Social Psychology*, 1964, 69, 418-426.

- Fiske, D. W., et al. Planning of research on effectiveness of psychotherapy. *American Psychologist*, 1970, 25, 725-737.
- Garfield, S. L., Prager, R. A., & Bergin, A. E. Evaluation of outcome in psychotherapy. *Journal of Consulting and Clinical Psychology*, 1971, 37, 307-313.
- Garwick, G., & Brintnall, J. (Eds.). *Proceedings of the second annual Goal Attainment Scaling Conference*. Minneapolis, Minn.: Program Evaluation Resource Center, 1974.
- Kiresuk, T. J., & Sherman, R. E. Goal Attainment Scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal*, 1968, 4, 443-453.
- Luborsky, L., Chandler, M., Auerbach, A. H., Cohen, J., & Bachrach, H. M. Factors influencing the outcome of psychotherapy: A review of quantitative research. *Psychological Bulletin*, 1971, 75, 145-185.
- Malan, D. H. The outcome problem in psychotherapy research: A historical review. *Archives of General Psychiatry*, 1973, 29, 719-729.
- Strupp, H. H., & Bloxom, A. L. Therapists' assessment of outcome. In I. E. Waskow & M. B. Parloff (Eds.), *Psychotherapy change measures*. Rockville, Md.: National Institute of Mental Health, 1975.
- Twain, D. Developing and implementing a research strategy. In E. L. Streuning & M. Guttentag (Eds.), *Handbook of evaluation research* (Vol. 1). Beverly Hills, Calif.: Sage Publication, 1975.
- Waskow, I. E., & Parloff, M. B. (Eds.). *Psychotherapy change measures*. Rockville, Md.: National Institute of Mental Health, 1975.
- Weiss, C. *Evaluation research*. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Received November 26, 1976 ■

Comparison of Cardiovascular Biofeedback, Neuromuscular Biofeedback, and Meditation in the Treatment of Borderline Essential Hypertension

Richard S. Surwit
Department of Psychiatry
Harvard Medical School

David Shapiro
Department of Psychiatry
University of California, Los Angeles

Michael I. Good
Department of Psychiatry, Harvard Medical School

This study compared the separate effects of three procedures for the reduction of high blood pressure (BP) in three treatment groups of eight patients each with medically verified borderline hypertension: (a) Biofeedback for simultaneous reductions in systolic BP and heart rate was aimed directly at reductions in BP. (b) Biofeedback for reductions in integrated forearm and frontalis muscle electromyographic activity was aimed at general muscular relaxation. (c) Meditation relaxation based on the "relaxation response" procedure developed by Herbert Benson was aimed at total bodily and "mental" relaxation. Each patient was studied in two baseline sessions, eight training sessions, and a 6-week follow-up. Half of the sample returned for a 1-year follow-up. Analysis of variance of the three treatment groups over eight training sessions, 20 trials per session, revealed significant effects for trials within sessions. However, there were no significant main effects or interactions related to differences between the treatment conditions or to changes in BP over the course of training sessions. Although all groups showed moderate reductions in BP as compared to initial values, no technique could be seen to produce a reduction in pressure greater than that observed in the baseline sessions. BPs of patients reporting for the 1-year follow-up were not different from pretreatment baseline levels.

Essential hypertension has long been of interest to those concerned with behavioral factors in disease. The disorder occurs in about 10% of the population and is known to be affected by behavioral therapies (Shapiro, Mainardi, & Surwit, 1977). It remains unclear, however, as to which behavioral approaches

can add most substantially to treatment programs for the disorder. The main purpose of this study was to compare the efficacy of three behavioral methods in the reduction of blood pressure (BP) in patients with essential hypertension.

Of the many behavioral techniques that have been used for treatment of hypertension, none has attracted as much recent interest as biofeedback training for the control of BP. Shapiro, Tursky, Gershon, and Stern (1969) developed a constant cuff method to monitor relative changes in BP occurring at each beat of the heart and to provide information to subjects about these changes. With this method, small but reliable relative increases and decreases in either systolic or diastolic BP were obtained in normal subjects (Shapiro, Schwartz, & Tursky, 1972; Shapiro et al., 1969). Larger decreases in BP were found when feed-

This research was supported by National Institute of Mental Health Research Grants MH-26923 and MH-25104, Training Grant MH-06415, and Office of Naval Research Contract N00014-75-C-0150, NR201-152. Computing assistance was obtained from the Health Sciences Computing Facility, University of California, Los Angeles, supported by the National Institutes of Health Special Research Resources Grant RR-3.

Requests for reprints should be sent to Richard S. Surwit, who is now at the Department of Psychiatry, Box 3416, Duke University Medical Center, Durham, North Carolina 27710.

back was provided for concomitant reductions in both heart rate (HR) and BP (Schwartz, 1972).

The constant cuff method of BP feedback was tested clinically in a series of case studies (Benson, Shapiro, Tursky, & Schwartz, 1971). Five patients with essential hypertension were trained to decrease systolic BP and showed reductions of 34, 29, 16, 16, and 17 mm of Hg after 33, 22, 34, 31, and 12 sessions of training, respectively. Using essentially the same procedure, Goldman, Kleinman, Snow, Bidus, and Korol (1975) reported decreases of 4% and 13% in systolic and diastolic BP, respectively, in seven patients with average baseline values of 167/109 mm of Hg. Kristt and Engel (1975) also reported success using the constant cuff feedback method. Patients suffering from essential hypertension and having a variety of cardiovascular complications were taught both to raise and to lower systolic BP and were able to achieve reductions in pressure averaging 10%-15% of their pretreatment baseline values.

Clinical data are less convincing for reductions in diastolic BP. Summarizing unpublished research, Miller (1975) noted that although some patients appeared capable of lowering diastolic BP, their pressure eventually drifted up over time. Negative results were also reported by Schwartz and Shapiro (1973). In this study, the constant cuff method was used to provide feedback for changes in diastolic BP. Only one out of seven patients studied showed a progressive reduction in diastolic BP; the others showed no change. However, Elder, Ruiz, Deabler, and Dillenkoffer (1973) reported using feedback and verbal praise to achieve a 20%-30% reduction in diastolic BP in patients who were not on antihypertensive medication. In addition, several studies previously discussed reported significant reductions in diastolic BP for patients who were actually trained to lower systolic pressure. Since morbidity has been shown to be related to elevations in systolic as well as in diastolic BP (Kannel, Gordon, & Schwartz, 1971), it may not be of great clinical consequence that diastolic pressure is more difficult to control with biofeedback methods.

Other biofeedback methods have been reported to be effective in the treatment of es-

sential hypertension. Moeller (1973) demonstrated that training patients to reduce frontalis electromyographic (EMG) activity through feedback training led to mean systolic and diastolic BP reductions of 13% in 36 essential hypertensive patients following 16 weeks of training. Patel (1973, 1975) used a combination of yoga and electrodermal (galvanic skin response; GSR) feedback to lower BP in hypertensive patients. In a well-controlled investigation, Patel and North (1975) randomly assigned 34 hypertensive patients to one of two treatment conditions. Patients in the first condition were initially taught relaxation with yoga and were then given GSR and EMG feedback in an effort to lower levels of autonomic and skeletal muscle activity. Patients in the second condition attended the same number of sessions (12), but they were told simply to recline on a lounge chair with no other specific instructions. Both groups showed reductions in BP, but the group that received yoga and biofeedback showed significantly greater reductions (from 168/100 to 141/88 mm of Hg) than the group that was simply instructed to rest (169/101 to 160/96 mm of Hg). Patients in the simple resting condition were then given 12 weeks of yoga plus biofeedback training and managed to lower their pressure to the level achieved by the first treatment group.

Another behavioral method reported to be effective in treating essential hypertension combines meditation and relaxation. Benson, Rosner, Marzetta, and Klemchuk (1974a, 1974b), using procedures derived from transcendental meditation (Maharishi Mahesh Yogi, 1966), demonstrated BP reductions in 22 untreated borderline and in 14 pharmacologically treated hypertensive patients. During the 6-week pretreatment baseline period, the borderline hypertension group showed an average systolic pressure of 147 mm of Hg and a diastolic pressure of 95 mm of Hg. During the 25 weeks of regular meditation relaxation, systolic BP was reduced by 8 mm of Hg and diastolic BP by 4 mm of Hg. The group being treated pharmacologically displayed an average systolic and diastolic BP of 146 mm of Hg and 92 mm of Hg, respectively, during a 5- to 6-week baseline period. The average reduction for this group during 20 weeks of regular meditation relaxation averaged 11 mm of Hg sys-

tolic and 5 mm of Hg diastolic. Using similar procedures, Stone and DeLeo (1974) recently compared meditation relaxation training to a no-treatment control in borderline hypertensive patients exhibiting pressures similar to those in the Benson et al. (1974a) study. Meditation relaxation was shown to produce small but significant pressure reductions (5–10 mm of Hg) as well as significant reductions in plasma dopamine-beta-hydroxylase activity and in furosemide-stimulated renin activity. On the basis of these data, Stone and DeLeo concluded that a reduction of peripheral adrenergic activity may be associated with the reduction of BP observed in the practice of meditation relaxation.

It seems clear that behavioral methods can produce reductions of BP. It is not clear, however, which methods are more effective than others and whether treatment effects are maintained consistently over time. Few follow-up data have been reported. Typically, various treatment approaches are combined, making it difficult to separate out critical variables and nonspecific or placebo effects. However, there is no easy solution to the problem of appropriate control procedures in evaluating behavioral therapies. The approach taken in this study was to compare several behavioral techniques. It was assumed that placebo effects would not differ across treatments, in that all treatments were "active" and could be expected to produce some change. Such a design allows a controlled comparison of efficacy and subject compliance, if subjects are randomly assigned to groups and treated over equal periods of time.

The present investigation compared three behavioral procedures in patients with essential hypertension: BP—biofeedback for reductions in BP and HR, EMG—biofeedback for simultaneous reductions in both frontalis and forearm muscle tension, and relax—a meditation relaxation practice. The rationale for the BP procedure was that BP biofeedback would facilitate direct control of reductions in BP and that simultaneous feedback for reduction in HR in this condition would serve to maximize the degree of BP reduction possible (Schwartz, 1972). The rationales for the other two conditions were also derived empirically. The EMG condition assumes that excessive

muscle activity is a critical factor in regulating high BP levels and that reduction in muscle activity facilitates pressure reductions.

By using the summed muscle activity of two muscle sites (frontalis and forearm) in the biofeedback procedure, it was assumed that total muscular relaxation would be facilitated. Further empirical support for muscle relaxation in essential hypertension has been reported by Shoemaker and Tasto (1975) with the use of a modified form of progressive relaxation. Moreover, it was further assumed that proprioceptive feedback of changes in muscle activity would aid in the development of learned control and maintenance of reduced muscle activity outside of the training situation. The relaxation procedure was assumed to maximize complete mental and physical relaxation. As hypothesized by Benson (1975), this method of relaxation produces a "hypometabolic state" that is presumed to represent a hypothalamic response that is antithetical to the "fight-flight" response and is consistent with a state of decreased sympathetic nervous system activity. According to Benson, the basic components of this form of relaxation are a mental device of mantra, a passive attitude, regular deep breathing, decreased muscle activity, and regular practice.

Follow-up evaluations of the treatment effects were made at 6 weeks and at 1 year after completion of the program.

Method

Subjects

Male and female volunteer subjects with hypertension were solicited through newspaper advertisements and physician referrals. After a preliminary telephone screening, potential subjects received (a) a letter of acknowledgement and an explanation of the study, (b) a confidential medical questionnaire focusing on history relevant to hypertension, and (c) a form requesting authorization for the volunteer's physician to release medical information. Subsequently, a description of the intent of the study was mailed to the volunteer's physician along with the authorized request for medical history including all available BP readings, summary of a physician's examination within the past year, and clinical laboratory data. Volunteers had to have their own physician agree to their participation in the study. The investigators did not assume the role of primary-care physicians.

Half of the volunteers considered for the study were taking antihypertensive or psychotropic medication,

Table 1
Characteristics of Patients

BP treatment group					EMG treatment group					Relax treatment group				
Age	Sex	L	M	BP*	Age	Sex	L	M	BP	Age	Sex	L	M	BP
52	Male	+	-	170/78	49	Male	+	+	140/90	42	Male	+	-	140/100
48	Male	+	-	138/90	57	Male	+	-	160/92	54	Male	-	+	150/90
42	Male	-	+	150/102	42	Male	-	-	150/95	43	Male	-	-	145/90
49	Male	-	+	158/94	49	Male	+	-	147/96	35	Male	+	+	151/91
52	Male	-	+	187/106	50	Female	-	+	158/103	27	Male	+	+	160/89
37	Female	+	-	149/103	46	Male	-	+	157/93	53	Male	+	+	153/96
37	Male	+	-	148/93	34	Male	+	+	159/87	51	Male	-	-	187/88
52	Female	-	+	159/90	59	Female	-	-	168/96	54	Female	-	-	165/95

Note. BP = blood pressure; L = lability; M = medication; EMG = electromyogram.

* Average prestudy blood pressure was measured in millimeters of mercury.

and the other half were not. The patients had a history of at least 1 year of labile or sustained systolic and/or diastolic hypertension (average systolic BP ≥ 140 , average diastolic BP ≥ 90). They were all less than 60 years old, and none had evident organic etiology for their hypertension, major complications related to the disease, or other serious illnesses (see Table 1).

Potential subjects whose histories met these criteria were then seen by one of the investigators for a detailed medical and psychosocial history, physical examination, and bilateral sitting and recumbent BP measurements. Subjects still meeting the criteria of the study were assigned to one of the three treatment groups. The groups were approximately matched for age, sex, current use or nonuse of relevant medication, type of blood pressure elevation (systolic and/or diastolic), and lability or stability of hypertension. BP was considered to be labile if the subject had a previous history of normal systolic and diastolic readings bracketed by elevated readings, with the lability being unrelated to medication effects. Subjects were requested to continue their current dietary and medication practices and to inform the investigators of any changes in their regimen during the study.

The BP and EMG groups were initially assigned nine members each, and the relax group, eight members. However, one subject in the BP group dropped out of the study, and one subject in the EMG group was excluded because of persistent tachycardia, leaving eight subjects in each group (Table 1).

Procedure

Subjects were scheduled to attend 10 sessions twice weekly for 5 weeks plus follow-up sessions approximately 6 weeks and 1 year after the 10th session. The first 2 sessions, 1 hour each, were used to obtain pre-treatment baseline measurements, and the remaining 8 1-1½ hours each, were experimental treatment sessions. To facilitate scheduling, approximately half of the subjects participated in either of two successive 5-week periods. In the baseline sessions, subjects were instructed to sit quietly and relax. The 3rd session marked the beginning of training. Subjects were read instruc-

tions appropriate for the experimental condition to which they were assigned. Instructions consisted of a brief reminder of the concept of psychophysiological BP reduction and a description of the general principle specific for the given treatment group. Instructions for the feedback or relaxation methods followed. The feedback groups were encouraged to use whatever mental strategy would result in the correct feedback. For the next 2 sessions, the instructions were paraphrased; and for the subsequent sessions, the subject was told: "I think you know the procedure by now. Are there any questions?"

Subjects were seated in a semirecumbent lounge chair in a wood-paneled sound-attenuated chamber. The following devices were connected to the subjects: a Biofeedback Systems, Inc., frontalis EMG headband consisting of three stainless-steel electrodes mounted in a rubber strap; two Beckman biopotential electrodes placed over the extensor for right forearm EMG; two silver-plate electrodes (right arm to left leg) and one silver-plate ground electrode on the right arm for the EKG; a respiration strain-gauge belt around the diaphragm; a crystal Korotkoff (K) sound microphone positioned over the left brachial artery and held in place by a BP cuff; and a Meditron electronic stethoscope affixed distal to the K sound microphone. During follow-up sessions, only the BP cuff and electronic stethoscope were attached. All recording was done in an adjacent instrument room. Two-way communication with the subject was maintained via an open intercom.

At the beginning of each session, with the subject alone in the training room, three BP measurements using the standard Riva-Rocci method were taken remotely by the experimenter in the adjoining instrument room. These readings were averaged to approximate the median systolic pressure, which was used as the initial constant cuff pressure setting. BP was then tracked continuously using the constant cuff method (Tursky, Shapiro, & Schwartz, 1972). All subjects were given 20 60-sec inflations (trials) separated by 30-sec rests. At the end of each trial, the cuff pressure was raised 2 mm of Hg if a subject showed K sounds following more than 75% of the heartbeats. Conversely, the cuff pressure was lowered by 2 mm of Hg for the next trial if K sounds followed less than 25% of the heartbeats. For each

group, sessions consisted of 20 60-sec trials, with the BP cuff inflated to median systolic pressure alternating with 30-sec intertrial intervals with the cuff deflated. In the BP and EMG training sessions, feedback was given only during the 20 60-sec trials. The constant cuff pressure on each trial constituted the basic BP data for all subjects.

For the BP group, BP and HR were measured, and changes in both were fed back using a pattern-feedback method (Schwartz, Shapiro, & Tursky, 1971). Briefly, BP was tracked as described previously. HR was tracked using a Lexington Instrument Company cardiographometer and a Grason-Stadler level detector. During the initial calibration period, median HR was determined. This is the HR at which half the interbeat intervals are faster and half are slower than that rate for a given series of heartbeats. The level detector was set at the median HR for the first trial and reset from trial to trial, depending on the number of interbeat intervals above and below that rate in the trial. The level was raised by 2 beats per minute (bpm) if 75% or more of the interbeat intervals were faster and lowered by 2 bpm if 25% or less of the intervals were faster than the median HR for the previous trial. In the feedback procedure, reductions of 2 bpm or more from the subject's median HR triggered the detector. During training, subjects in the BP group received feedback whenever HR dropped below the median level accompanied by a simultaneous reduction of BP (absence of K sound within 400 msec following the R wave). Feedback consisted of an auditory signal and visual feedback (cumulative digital meter). Subjects were instructed as follows:

During periods when the cuff is inflated, feedback in the form of a tone and increment on the meter will be provided whenever the equipment detects a decrease in your BP. If your BP goes up, or stays the same, no feedback will be given. Your job is to get as much feedback as possible (a high meter count) on each trial.

The EMG group heard a change in click signal frequency and observed the deflection of an analogue meter as average EMG activity varied during each 60-sec trial. The right forearm extensor and frontalis EMG readings were integrated by two Beckman EMG couplers. The combined average of these signals was used to drive a voltage-controlled oscillator, which produced auditory clicks directly proportional to integrated EMG activity. The integrated average of the two EMG signals also drove a large analogue meter placed directly in front of the subject. The task of the EMG group was to lower EMG activity using both auditory and visual feedback. Briefly, instructions were as follows:

Your task is to reduce your blood pressure by relaxing your muscles. During periods when the cuff is inflated, feedback for your muscle activity will be provided. As your muscle activity decreases, the meter in front of you will move to the left, and the clicking sound that you will hear will decrease in frequency. Conversely, if you tense your muscles, the meter will move toward the right, and the clicks

will increase in frequency. Your job is to decrease the click rate and keep the meter over to the left as much as possible.

The relax group was asked to meditate following the relaxation response method of Benson (1975) during the 20 trials of BP measurement. Briefly, instructions were as follows:

During the session, the BP cuff will be inflated from time to time for 1-minute periods. No other stimuli will be provided. When you are told to start, close your eyes and relax your muscles. Breathe through your nose. As you breathe out, say the word *one* silently to yourself. For example, breathe in...out, one, in...out, one, etc. Breathe easily and naturally. Continue until you are told to stop. Do not worry whether you are successful in achieving a deep level of relaxation. Maintain a passive attitude. If distracting thoughts occur, try to ignore them by not dwelling upon them and return to repeating one.

At the end of each session, subjects in each group were shown a cumulative graph of their total progress in lowering BP.

After the first training session, subjects were instructed to practice lowering BP as often as possible between sessions using whatever strategy they had arrived at during the laboratory sessions but without the aid of a sphygmomanometer or other instrumentation. Subjects were requested to record the nature and time of this practice, and the record sheets were collected before beginning each of the remaining seven training sessions. Generally subjects reported no significant changes during the course of training in their compliance with home practice. Nor were there any reported significant changes in medication during the course of training.

After the eighth training session, subjects were asked to return to the laboratory for a 6-week and a 1-year follow-up session. All subjects returned for the 6-week follow up, but only 13 returned for the 1-year follow-up (7 BP, 3 EMG, 3 relax). During both follow-up sessions, a BP cuff and an electronic stethoscope were attached to the left arm. The subjects were then left alone in the laboratory, and they were asked to practice the strategy for lowering pressure that they had learned during training. No feedback was provided for any subjects during these sessions.

Two female experimenters ran equal numbers of subjects in all three groups. Each subject was run by the same experimenter throughout the experiment.

Results

Pretreatment Baseline Data

To determine if the three treatment groups differed in average BP prior to the beginning of training, an analysis was made of five pretreatment pressures: (a) average BP obtained in the prior medical history, (b) average BP obtained

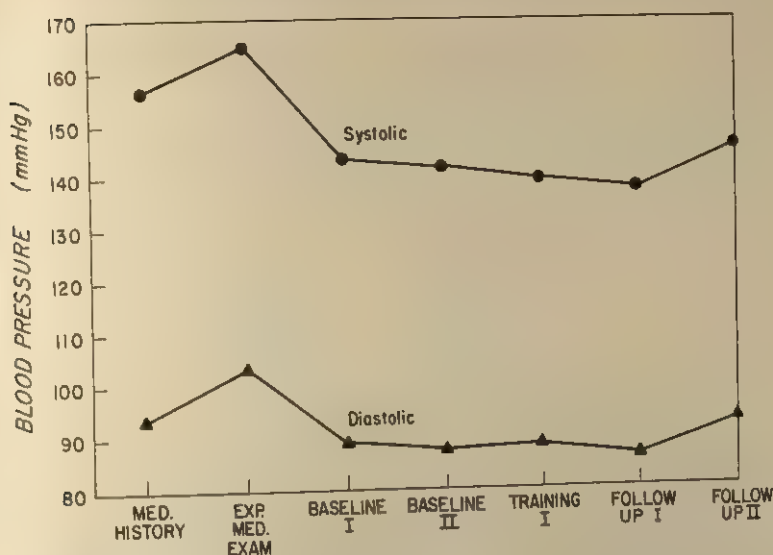


Figure 1. Mean blood pressure (measured in millimeters of mercury; mm Hg) collapsed over treatment groups over the course of the study. (For Follow-up II, $n = 13$, Med. = medical; Exp. = experimental.)

in the physician's examination, (c) average BP at the beginning of Baseline Session 1, (d) average BP at the beginning of Baseline Session 2, and (e) average BP at the beginning of Training Session 1. A two-way analysis of variance was carried out for the three treatment groups and five repeated measures of BP in each group. BMD 08V and BMD P2V programs were used in this analysis and in subsequent statistical analyses (University of California, Los Angeles, Health Sciences Computing Facility). Results of the analysis of variance showed that baseline differences between treatment groups were not significant for either systolic or diastolic pressure. The three groups had comparable BP values prior to the beginning of active treatment. However, pressure varied considerably under the different conditions of observation. Pressures were highest for the physician's examination and previous medical history and lowest for the two baseline sessions and the initial training session (Figure 1). A reduction of 5 mm of Hg in systolic pressure and a slight decrease (1 mm of Hg) in diastolic pressure were observed from Baseline Session 1 to Training Session 1. Differences among the five baseline measurements were significant for both systolic and diastolic values, $F(4, 84) = 36.79$, $p < .01$; $F(4, 84) = 30.35$, $p < .01$. Systolic blood pressure

averaged over all 24 patients differed by 17 mm of Hg between medical history value (156 mm of Hg) and Training Session 1 (139 mm of Hg); diastolic blood pressure was lower by 11 mm of Hg (94 mm of Hg and 83 mm of Hg, respectively). Reductions of still larger magnitudes (26/15 mm of Hg) were found when the project physician's recordings were compared to the first training session pressures. These reductions were much larger than any observed during the course of the actual treatment sessions. These data indicate the potency of environmental influences on BP. They also point out the issue of proper baseline determinations and habituation as factors to be accounted for in evaluating the effects of treatment programs of this kind.

Treatment Data

Average systolic pressures for each training session are shown in Figure 2. Also shown are the average systolic pressures for the baseline days and Follow-up Day 1. The graph suggests that the relax group decreased in pressure and the other two groups increased in pressure, comparing the baseline days and Training Day 1. Several preliminary analyses of variance were computed to examine this apparent differential trend. First, an analysis of variance

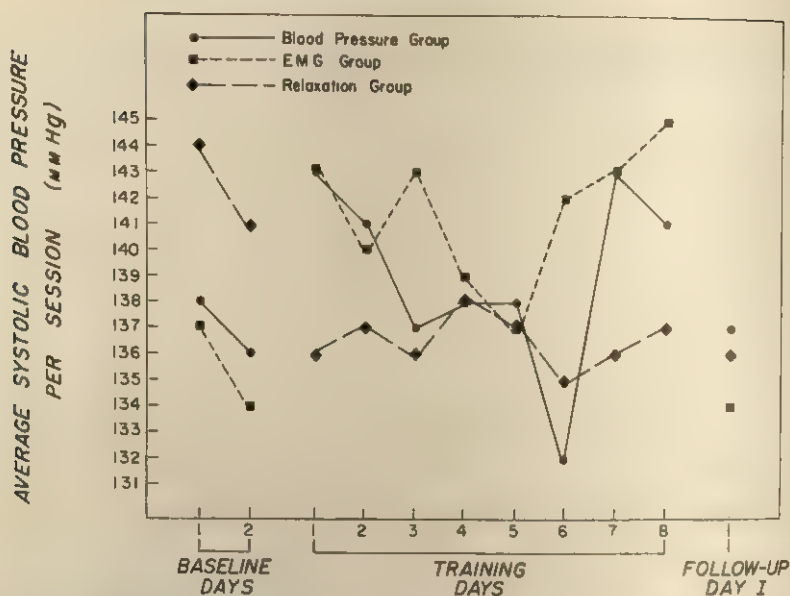


Figure 2. Mean blood pressure (measured in millimeters of mercury; mm Hg) per session in the three treatment groups. (EMG = electromyogram.)

was carried out for the three treatment groups, comparing Baseline Session 2 and Training Session 1. Neither the interaction between treatment and session nor the session difference were significant. Second, the decrease in systolic BP for the relax group by itself was not statistically significant. Third, an analysis of covariance for the values during the 8 training days adjusting for differences in Baseline Day 2 did not yield a significant treatment effect or a significant Treatment \times Session interaction. Therefore, the treatment data were analyzed without adjusting for baseline values, which were not significantly different between treatment conditions.

The primary treatment data consisted of the systolic pressures obtained during eight sessions of training, 20 values per session, with each value being the median systolic pressure derived from the constant cuff method for each 60-sec trial period. An analysis of variance was carried out for the three treatment groups with repeated measures for eight sessions and 20 trials within each session. A significant main effect was obtained for trials, $F(19, 399) = 15.72$, $p < .01$. On the average, pooling sessions and groups, the initial value in a session was 140.2 mm of Hg, and the 20th value was 136.8 mm of Hg. No other main effects or interactions were statistically significant.

Although the overall analysis of variance for training data indicated no significant effects between groups, inspection of the curves suggested that each treatment group showed a somewhat different pattern of apparent change or habituation over training sessions. For purposes of description only, an analysis of variance for trends was used to examine change over the course of treatment in each group separately. Only linear and quadratic trends were considered. Of the three treatment conditions only BP showed a significant main effect for sessions, $F(7, 49) = 2.52$, $p < .03$. The quadratic effect for sessions was also significant in this group, $F(1, 7) = 24.87$, $p < .002$. In accord with results presented earlier, each group showed a highly significant main effect for trials. Significant linear trends for trials occurred in EMG and relax, considering all sessions or the first six sessions only. The linear trial trend was not significant for the BP group; however, the quadratic trend for trials was significant in this group for all sessions.

Table 2 gives the average 1st and 20th pressure value over Baseline, Training, and Follow-up 1 days. As can be seen in the table, reductions in pressure over the 20 trials within a session were larger in baseline sessions than in training sessions. An analysis of variance comparing Baseline Session 2 with Training Session

Table 2
Systolic Blood Pressure: First and Last Trial Values

Session		Treatment group			All subjects
		BP	EMG	Relax	
Baseline	1	141-136	142-134	142-142	142-138
	2	138-134	140-132	145-136	141-134
Training	1	143-142	142-140	137-136	141-139
	2	140-139	141-137	138-134	140-137
	3	138-137	145-140	138-133	140-137
	4	137-137	138-137	140-138	138-137
	5	137-136	141-134	139-134	139-135
	6	134-128	141-140	138-132	138-133
	7	143-142	143-140	139-133	142-138
	8	143-139	146-143	139-133	143-138
Follow-up	1	139-135	136-130	137-134	138-133

Note. BP = blood pressure, measured in millimeters of mercury; EMG = electromyogram. $n = 8$ for each group.

sion 6 (a session showing a good effect) supports this conclusion ($p < .01$). In Baseline 2, systolic pressure started at 141 mm of Hg in Trial 1 and ended at 134 mm of Hg in Trial 20. Comparable values for Session 6 were 138 mm of Hg and 133 mm of Hg.

Table 2 also reveals the other patterns discussed above. The relax data appeared remarkably uniform from session to session, with Trial 1 values varying from 137 mm of Hg to 140 mm of Hg to 139 mm of Hg over sessions and with reductions in pressure over trials varying from 1 mm of Hg to 6 mm of Hg. The BP group showed little decrease over trials, with the exception of Session 6, and the previously noted trend toward reduction over Sessions 1-6, from 143 mm of Hg to 134 mm of Hg for Trial 1 and from 142 mm of Hg to 128 mm of Hg for Trial 20. Findings in the EMG group were less consistent. In this group, Trial 1 values varied up and down, and reductions during sessions varied from 1 mm of Hg to 7 mm of Hg.

Finally, the treatment session data were re-analyzed for the effects of individual differences in lability and medication. Two separate analyses of variance were carried out, dividing each treatment group into two equal subgroups on the basis of lability and medication (Table 1).

In the lability analysis of variance, a significant three-way interaction occurred for Session \times Trials \times Lability, $F(133, 2394) = 1.27$, $p < .02$. Although this interaction is difficult to interpret, it appears from the data

that labile patients showed more variable, including occasional larger, reductions in pressure over trials and over sessions than did nonlabile patients. These differences were not related to treatment condition. As for medication, three significant interactions were obtained: (a) Trials \times Medication, $F(19, 342) = 1.83$, $p < .02$. On the average, for all treatments, medicated patients showed a slightly greater reduction (about 2 mm of Hg) in systolic pressure from Trial 1 to Trial 20 than did nonmedicated patients. (b) Session \times Treatment \times Medication, $F(14, 126) = 1.92$, $p < .03$. Patterns of change over sessions differed for the three groups according to medication, but no clear interpretation can be arrived at. (c) Session \times Trials \times Treatment \times Medication, $F(266, 2394) = 1.38$, $p < .01$. A four-way interaction such as this can only be speculated about. Examining the first and last trials for each treatment group broken down according to medication, we find that in general non-medicated patients in the BP group showed greater within-session reductions in pressure and increases in these reductions over sessions than medicated patients. In EMG, medicated patients seemed to do consistently better over sessions in terms of pressure reduction. In relax, the trends were about the same in both medicated and nonmedicated patients.

Associated Changes During Treatment

Data on HR in beats per minute were available on each trial during the training sessions.

Table 3
Physiological Changes on Training Day 6

Measure	Treatment group		
	BP	EMG	Relax
BP	-6.0	-1.0	-6.0
Forearm EMG (μ v)	-4.4	-4.4	-10.6
Frontalis EMG (μ v)	+4.4	-12.2	-23.1

Note. BP = blood pressure, measured in millimeters of mercury; EMG = electromyogram.

An analysis of variance comparable to the one carried out on systolic pressure was done. No significant effects related to treatment group were found. As in the case of systolic blood pressure, a significant main effect for trials was observed, $F(19, 399) = 115.33$, $p < .001$. On the average, there was a 3-bpm decrease in HR during sessions from Trial 1 to Trial 20. Separate analyses of each group did not reveal any significant trends over sessions for any group.

Integrated EMG was recorded from both frontalis and right forearm muscles during all baseline and training sessions. An analysis of the change in EMG from Trial 1 to Trial 20 was carried out on Training Session 6 data. Although mean systolic BP of the three groups varied on that day (Figure 2), the change in systolic BP from Trial 1 to Trial 20 was not significant.

Table 3 gives the average systolic BP change, average forearm extensor EMG change, and average frontalis EMG change from Trials 1 to 20 on Day 6. No significant differences between the groups were found in forearm EMG activity. A Link-Wallace shortcut analysis of variance (Mosteller & Bush, 1954) revealed a significant difference for frontalis EMG activity between the BP and relax groups ($p < .05$). Differences in frontalis EMG activity between BP and EMG and EMG and relax were not significant. These data suggest that comparable reductions in BP can be achieved with differing skeletal muscular responses.

Follow-up Data

Turning to the 6-week follow-up data, no significant differences were obtained between groups. Average systolic and diastolic values were quite comparable (Figure 2), and the first and last trial values were also very similar

(Table 2) from group to group. Comparing these values to the medical history or to initial value obtained by the physician, the reductions in pressure for all subjects were 18/8 mm of Hg and 27/17 mm of Hg, respectively. However, comparing these same follow-up values to the initial values obtained at the beginning of the first training session (actual third pretraining visit of patients to the laboratory), the reductions were quite small—6 mm of Hg systolic and 3 mm of Hg diastolic. Whatever benefit there may have been during the training itself, it appears to be of a small magnitude. It is clear that significant reductions in pressure had already occurred prior to the actual inception of treatment, probably as a result of habituation to the laboratory situation (see Figure 1). The apparent reduction in pressure for the relax group at the time of the 6-week follow-up as compared with the baseline days (see Figure 2) was not statistically significant.

In the 1-year follow-up, only about half of the total sample of patients returned for evaluation: 7 BP, 3 EMG, and 3 relax. At this time, the initial in-laboratory average pressures for the three groups were, respectively, 138/88 mm of Hg, 148/92 mm of Hg, and 138/91 mm of Hg. Given the small samples, it is not possible to determine whether the groups differed significantly at this time. The average pressure for the 13 patients was 141/90 mm of Hg, about 3/3 mm of Hg higher than the level obtained at the 6-week follow-up and at the beginning of Training Session 1 (Figure 1).

Discussion

The purpose of this study was to compare the effectiveness of three different, purely behavioral, means of lowering blood pressure in patients with essential hypertension. It was expected on the basis of previous studies that the procedures would be effective to some degree. This was not the case. There was little evidence, either during the course of the treatment sessions or in the follow-up evaluations, of significant reductions in pressure. Considering the initial pressure values obtained immediately prior to the beginning of training as a baseline, the average reduction for all 24 patients at the time of the 6-week follow-up was 1 mm of Hg systolic and 2 mm of Hg diastolic. On the basis

of data available from about half the sample at the time of the 1-year follow-up, pressures were back to pretreatment levels. No one treatment procedure was obviously better than any other. Within sessions, they all resulted in significant reduction of pressure of about 4 mm of Hg systolic on the average.

For each of the methods used in this study, several earlier studies have been cited that reported average reductions in pressure ranging from 5% to 15% of pretraining baseline values. The failure to obtain even small reductions in the present study is therefore quite puzzling. Some speculations can be offered as the reasons for this failure and the apparent inconsistency between previous and present findings. Consider the BP group findings. In this group, there was only a small reduction in systolic pressure over the 20 trials within any one session, typically 1 mm of Hg. Benson et al. (1971) reported average within-session reductions of 4.8 mm of Hg in seven patients, and this value is in accord with the change observed in normotensive subjects (Shapiro et al., 1969). Both of the latter studies contained simple systolic pressure feedback. Since we know that patients can respond positively to systolic feedback by itself, it may be that the additional task of training to reduce HR at the same time made it more difficult for these patients to reduce BP. The average HR in these BP patients was not elevated, possibly further adding to the difficulty. Previous studies involving behavioral treatment of hypertension through direct BP biofeedback used patients with sustained hypertension whose pressures were usually higher than subjects studied here. Subjects in the Benson et al. (1971) study showed average pretraining systolic pressures of 165 mm of Hg, and Goldman et al. (1975) reported a sample average of 167 mm of Hg, whereas our subjects showed average pretraining pressures of 141 mm of Hg during the first three laboratory exposures prior to training. It is possible that the effectiveness of behavioral treatment follows the law of initial values (Wilder, 1957), and that only patients with relatively elevated pressures will show substantial benefit from such treatment. Furthermore, although the average pressures for our sample of patients were relatively high as recorded in various medical examinations prior to treatment, the values were

considerably reduced in the quiet, relatively nondemanding laboratory conditions. The usefulness of biofeedback training or of other methods of behavioral control under conditions in which the symptom is not shown in full force may be questionable.

Another difference between this study and previous reports is that the present investigation gave all subjects a rather limited amount of training and attention. Only 8 training sessions were used in the present study as compared to up to 34 sessions in the Benson et al. (1971) study or 42 sessions in the Kristt and Engel (1975) study. It is of interest, however, that the one subject studied by Benson et al. (1971) who had both similar baseline pressures and number of training sessions showed a reduction in pressure similar to that of subjects in our study. Also, in the Goldman et al. (1975) study in which only nine treatment sessions were used, systolic reductions of only 4% were observed although diastolic changes were much larger.

As for the EMG group, we asked these patients to lower muscle activity in two sites, forearm and forehead. This double requirement may have contributed to the apparent difficulty in achieving total muscular relaxation. In any case, the EMG group data are quite variable and inconsistent with earlier reports of success. We could speculate that some other more potent ingredient such as the suggestion to relax in the treatment process or an expectancy of success was present in the other studies and missing in ours.

In the relax group, the pressure data were remarkably consistent from day-to-day treatment and relatively consistent within sessions. It seems that the relax group involves the least task or problem-solving orientation of the three procedures. This is reflected in the apparent immediate reduction in pressure at the very beginning of Training Session 1 and the consistent 4-5 mm of Hg reduction in pressure from beginning to end of each session. The lack of pressure reactivity reflects the passive, relaxed stance involved in the procedure, and the data attest to the ability of patients to comply with the instructions that seem to enhance their ability to relax and to show rapid habituation to the treatment situation. If there is any lasting therapeutic benefit to such a pro-

cess, it was not revealed in any session-by-session trend or in the follow-up data. Because of its inherent simplicity, however, the meditation relaxation procedure could be seen to have an advantage over other methods.

These results, in light of the interpretations and possible explanations offered here, suggest possible refinements in future research. First, we need more systematically to segregate labile from fixed patients, to account more thoroughly for the effects of medication, and to make a more thorough assessment of cardiovascular status to determine the suitability of feedback of different pressure (diastolic, systolic) or cardiac (heart rate, cardiac output) indices, whether singly or in patterns. Second, the arbitrary use of a fixed number of training sessions may not be adequate, and more than eight sessions may be required. Third, the design of treatment procedures must carefully consider the degree to which task involvement should play a part in the process. For some patients, simple passive, nontask conditions, such as mediation relaxation, may be appropriate. In others for whom abnormal reactivity to stressful situations or other eliciting events is a factor, the learning process could well incorporate a task, such as in biofeedback for pressure reductions, or a task involving alterations of physiological response in coping with stressful stimuli (Sirota, Schwartz, & Shapiro, 1974).

Finally, rather large differences were observed between pressure levels obtained in the medical history and in our physician's examination as compared with the levels obtained in quiet laboratory conditions, particularly after a few baseline nontask recording sessions. These differences range up to 20 mm of Hg systolic and 15 mm of Hg diastolic. They attest to the potency of real-life variables (e.g., doctor's examination) versus the lack of potency of nonstimulating laboratory conditions. Such variability has been noted repeatedly in the literature (Julius & Schork, 1971). Engel (Note 1) has suggested that the physician effect or habituation to the physician and medical situation may account in part for the placebo effect noted in hypertensive research (Miller, 1974). The commonly reported reductions in pressure in many behavioral treatment studies may also reflect this process of habituation. It is obvious that we need to pay more attention to the

sources of variability in blood pressure, particularly in patients with essential hypertension, as these sources may tell us about the nature of the disorder and give us clues about the design of relevant behavioral treatment procedures and appropriate methods of evaluating treatment effects.

Reference Note

1. Engel, B. T. Personal communications, February 27 to March 2, 1976.

References

- Benson, H. *The relaxation response*. New York: Morrow, 1975.
- Benson, H., Rosner, B. A., Marzetta, B. R., & Klemchuk, H. M. Decreased blood pressure in borderline hypertensive subjects who practiced meditation. *Journal of Chronic Diseases*, 1974, 27, 163-169. (a)
- Benson, H., Rosner, B. A., Marzetta, B. R., & Klemchuk, H. M. Decreased blood-pressure in pharmacologically treated hypertensive patients who regularly elicited the relaxation response. *Lancet*, 1974, 1, 289-291. (b)
- Benson, H., Shapiro, D., Tursky, B., & Schwartz, G. E. Decreased systolic blood pressure through operant conditioning techniques in patients with essential hypertension. *Science*, 1971, 173, 740-742.
- Elder, S. T., Ruiz, Z. R., Deabler, H. L., & Dillenkoffer, R. L. Instrumental conditioning of diastolic blood pressure in essential hypertensive patients. *Journal of Applied Behavior Analysis*, 1973, 6, 377-382.
- Goldman, H., Kleinman, K. M., Snow, M. Y., Bidus, D. R., & Korol, B. Relationship between essential hypertension and cognitive functioning: Effects of biofeedback. *Psychophysiology*, 1975, 12, 569-573.
- Julius, S., & Schork, M. A. Borderline hypertension—A critical review. *Journal of Chronic Diseases*, 1971, 23, 723-754.
- Kannel, W. B., Gordon, T., & Schwartz, M. J. Systolic versus diastolic blood pressure and risk of coronary heart disease. *American Journal of Cardiology*, 1971, 27, 335-343.
- Kristt, D. A., & Engel, B. T. Learned control of blood pressure in patients with high blood pressure. *Circulation*, 1975, 51, 370-378.
- Maharishi Mahesh Yogi. *The science of being and the art of living*. London: International SRM Publications, 1966.
- Miller, N. E. Preface. In N. E. Miller, T. X. Barber, L. V. DiCara, J. Kamiya, D. Shapiro, & J. Stoyva (Eds.), *Biofeedback and self-control 1973: An Aldine annual on the regulation of bodily processes and consciousness*. Chicago: Aldine, 1974.
- Miller, N. E. Applications of learning and biofeedback to psychiatry and medicine. In A. M. Freedman, H. I. Kaplan, & B. J. Sadock (Eds.), *Comprehensive textbook of psychiatry-II*. Baltimore, Md.: Williams & Wilkins, 1975.
- Moeller, T. A. *Reduction of arterial blood pressure through relaxation training and correlates of personality in hy-*

- peritives*. Unpublished doctoral dissertation, Nova University, Fort Lauderdale, Florida, 1973.
- Mosteller, F., & Bush, R. R. Selected quantitative techniques, In G. Lindzey (Ed.), *Handbook of social psychology* (Vol 1). Cambridge, Mass.: Addison-Wesley, 1954.
- Patel, C. H. Yoga and biofeedback in the management of hypertension. *Lancet*, 1973, 1, 1053-1055.
- Patel, C. 12-month follow-up of yoga and biofeedback in the management of hypertension. *Lancet*, 1975, 1, 62-64.
- Patel, C. H., & North, W. R. S. Randomized controlled trial of yoga and biofeedback in management of hypertension. *Lancet*, 1975, 1, 93-95.
- Schwartz, G. E. Voluntary control of human cardiovascular integration and differentiation through feedback and reward. *Science*, 1972, 175, 90-93.
- Schwartz, G. E., & Shapiro, D. Biofeedback and essential hypertension: Current findings and theoretical concerns. *Seminars in Psychiatry*, 1973, 5, 493-503.
- Schwartz, G. E., Shapiro, D., & Tursky, B. Learned control of cardiovascular integration in man through operant conditioning. *Psychosomatic Medicine*, 1971, 33, 57-62.
- Shapiro, D., Mainardi, J. A., & Surwit, R. S. Biofeedback and self-regulation in essential hypertension. In G. E. Schwartz & J. Beatty (Eds.), *Biofeedback: Theory and research*. New York: Academic Press, 1977.
- Shapiro, D., Schwartz, G. E., & Tursky, B. Control of diastolic blood pressure in man by feedback and reinforcement. *Psychophysiology*, 1972, 9, 296-304.
- Shapiro, D., Tursky, B., Gershon, E., & Stern, M. Effects of feedback and reinforcement on the control of human systolic blood pressure. *Science*, 1969, 163, 588-590.
- Shoemaker, J. E., & Tasto, D. L. The effects of muscle relaxation on blood pressure of essential hypertensives. *Behaviour Research and Therapy*, 1975, 13, 29-43.
- Sirota, A. D., Schwartz, G. E., & Shapiro, D. Voluntary control of human heart rate: Effect on reaction to aversive stimulation. *Journal of Abnormal Psychology*, 1974, 83, 261-267.
- Stone, R. A., & DeLeo, J. Psychotherapeutic control of hypertension. *New England Journal of Medicine*, 1974, 294, 80-84.
- Tursky, B., Shapiro, D., & Schwartz, G. E. Automated constant cuff-pressure system to measure average systolic and diastolic blood pressure in man. *IEEE Transactions on Biomedical Engineering*, 1972, 19, 271-276.
- Wilder, J. The law of initial values in neurology and psychiatry. Facts and problems. *Journal of Nervous and Mental Disease*, 1957, 125, 73-86.

Received March 28, 1977 ■

Effects of Transcendental Meditation and Muscle Relaxation on Trait Anxiety, Maladjustment, Locus of Control, and Drug Use

David C. Zuroff and J. Conrad Schwarz
University of Connecticut

Sixty undergraduate volunteers were randomly assigned to receive training in transcendental meditation (TM), training in a muscle relaxation technique, or no treatment. The training in muscle relaxation was designed to be maximally similar in structure and atmosphere to training in TM. Measures of trait anxiety, locus of control, maladjustment, and drug use were collected before and after the 9-week treatment period. On a behavioral measure of trait anxiety, the scores of all three groups decreased equally, but on a self-report measure the TM subjects reported steady decreases in anxiety, whereas the scores of the other two groups remained unchanged. There were no differences in maladjustment, locus of control, or drug use as a function of treatment. Although TM subjects held higher expectancies for benefits, and were slightly more regular in practicing their technique, individual differences in expectancy and frequency of practice were not correlated with degree of reported anxiety reduction. It is concluded that TM may reduce trait anxiety, but it has not been shown to be of value in inducing general personality change.

Studies of the effectiveness of transcendental meditation (TM) have generally lacked the methodological rigor and sophistication that is now expected in studies of psychotherapy outcome (Smith, 1975). Three particularly pervasive flaws have been the following: the failure to obtain initially equivalent treatment and control groups by random assignment of subjects, the tendency to rely exclusively on self-report measures, and the failure to use "placebo" treatments to control for nonspecific treatment effects. It is interesting to note that with the exception of Smith's (1976) methodologically sound study, previously published research has been uniformly favorable to TM.

This article is based on portions of a thesis submitted by the first author in partial fulfillment of the requirements for the MA degree at the University of Connecticut, 1975. This investigation was supported by grants from the University of Connecticut Research Foundation, the University Computer Center, and Research Grant GS-31020X from the National Science Foundation.

Requests for reprints should be sent to J. Conrad Schwarz, Department of Psychology, Box U-20, University of Connecticut, Storrs, Connecticut 06268.

In particular, changes in the "healthier" direction have been reported for each of the variables of interest in the present study: anxiety (Ballou, 1973; Ferguson & Gowan, 1973; Hjelle, 1974; Orme-Johnson, Arthur, Franklin, O'Connell, & Zold, 1973); maladjustment (Ferguson & Gowan, 1973; Hjelle, 1974; Seeman, Nidich, & Banta, 1972); locus of control (Hjelle, 1974); and marijuana use (Shafii, Lavelly, & Jaffee, 1974).

These results must be interpreted in light of the design flaws mentioned above; none of these studies used a placebo control group, and only Ballou (1973) achieved a true experimental design. Based on his review of these and other studies, Smith (1975) rightly concluded that methodological flaws in the studies reviewed were such that there "is not clear evidence that meditation is in and of itself therapeutic" (p. 562).

Smith (1976) attempted to overcome the limitations of these studies by randomly assigning students who had volunteered to obtain free treatment for reducing anxiety to receive either TM, no treatment, or a treatment designed to be equivalent to TM in

terms of both expectations for benefits and the repeated quiet sitting involved in meditation. Using the Spielberger State-Trait Anxiety Inventory (Spielberger, 1972), the Tennessee Self-Concept Scale, and the Sixteen Personality Factor Questionnaire, Smith found that on virtually every measure of psychopathology, the two treatment groups improved significantly more than the untreated controls. However, the TM subjects showed no more improvement than subjects exposed to the placebo treatment. This finding was replicated in a second experiment comparing a TM-like meditation technique with an "antimeditation" technique involving the active generation of positive thoughts. Smith (1976) concluded that the crucial therapeutic agent in meditation is not the focusing of attention on the mantra (a Sanskrit word personally prescribed for each meditator) but rather some combination of just sitting and the expectation that the technique one is practicing is therapeutic.

Although Smith's study is weakened by the high dropout rate—59% for TM and 53% for the control treatment in the first study—and the reliance on self-report measures, it certainly casts considerable doubt on earlier claims that TM has specific treatment effects on anxiety and, more globally, on psychological adjustment.

The present study, like Smith's, provided for random assignment of subjects to the three experimental groups and for a nonspecific effects control group. However, the muscle relaxation technique was designed to control not only for the nonspecific effects of TM, including placebo effects and the effects of simply sitting regularly, but also for its demonstrated properties of autonomic arousal reduction (Wallace, 1970).

Method

Subjects

Subjects were recruited from the introductory psychology course at the University of Connecticut for "an experiment concerning transcendental meditation and a muscle relaxation technique." Students who had prior training in any kind of meditation were not eligible, so our sample may not be fully representative of those who independently seek instruction in TM. However, our subjects were interested in meditation, and in fact almost all subjects in the final sample stated that they had volunteered hoping to be assigned to the TM group.

One hundred fifty students volunteered, and from these we randomly selected 30 males and 31 females who were not receiving therapy or counseling: 1 female dropped out of the TM group shortly before training. All subjects were required to post a \$10 deposit, forfeitable to the scholarship fund if he/she failed to complete all the testing. Sixty-seven percent of the subjects were 18 years old; 75% were in their first or second semester; 66% were firstborn or only children; and 52% were Catholic by birth.

General Experimental Design

Subjects were randomly assigned to receive either training in TM, training in muscle relaxation, or no treatment. Measurements were made before and immediately after the training, at the midpoint of the treatment period, and again at the end of the 9-week treatment period.

Transcendental Meditation

Instruction in TM, which consists of two group lectures of about 1 hour each, 1 hour of individual instruction in the technique itself, and three additional 1-hour group meetings, was provided by experienced initiators from the Storrs branch of the Student International Meditation Society. Throughout the training period it is repeatedly stressed that TM is easy, natural, and spontaneous, and that regular practice will inevitably bring a wide range of benefits. The only departure from the standardized procedure introduced by the experiment was the use of grant funds to pay subjects' \$45 initiation fees.

Muscle Relaxation

The training procedure in muscle relaxation was designed to duplicate as nearly as possible the structure and atmosphere of TM instruction. Considerable attention was devoted to making the treatment equivalent to TM in credibility and attractiveness and to heightening subjects' expectations for benefits. The training in muscle relaxation followed the same sequence and provided the same amount of contact with instructors as the training in TM.

At the two initial lectures, the first author presented a rationale for the technique, which suggested that "tension" was the root cause of human unhappiness and that muscle relaxation would reduce both mental tension and, in the long run, chronic physical tension or "body armor." The experience of muscle relaxation was described as one of pleasant relaxation in which the mind floats off gently, returning refreshed and renewed.

Subjects were taught muscle relaxation by three advanced doctoral students in clinical psychology (two males, one female) and a female doctoral student in Family Life Education who held a master's degree in clinical psychology.¹ They were all experienced in teach-

¹ We would like to thank John Dufresne, Ron Lajoy, Margaret Nichols, and Ruta Teisman for their work in teaching the muscle relaxation subjects.

ing the technique and were supervised by a licensed clinical psychologist. They followed Paul's (1966) instructions for inducing muscle relaxation, spending between 20 and 40 minutes with each subject. Paul's (1966) procedure is an accelerated form of Jacobson's (1938) technique that involves alternately tensing and releasing muscle groups until a state of deep muscle relaxation is achieved. Subjects were instructed to practice at home for approximately 20 minutes twice per day.

The group meetings, which were held during the 3 evenings following the individual instruction, included discussion of problems encountered by the subjects, additional lecture material, and group practice of muscle relaxation. The lecture material consisted of scientific-sounding explanations of how muscle relaxation reduces muscle tension and promotes personality growth and stressed the simplicity, ease, and inevitable efficacy of the technique.

No Treatment Control

Subjects in the no-treatment control group were not given any training nor were they asked to alter their lives in any special way. They were exposed to virtually the same questionnaires and measurement procedures as subjects in the other groups.

Measures

Assessments were made in five areas: background and personality variables; expectancies; extent of arousal reduction; psychological maladjustment; and trait anxiety. Both behavioral and self-report measures of trait anxiety were used.

Background and personality variables. All subjects were administered Rotter's (1966) Locus of Control scale; the Social Desirability Scale (Crowne & Marlowe, 1960); and questionnaires concerned with demographic characteristics, previous and current drug and alcohol use, and experiences with TM and muscle relaxation.

Expectancies. The expectancy questionnaire included 25 possible benefits that muscle relaxation and TM subjects rated on an 11-point scale (0%-100%, in 10% intervals) for likelihood that regular practice would lead to that benefit. The ratings were summed to yield a total expectancy score. The items were drawn from the four general areas of improvements in subjects' emotional lives, achievements, interpersonal relations, and health. Some sample items are "lessened feelings of anxiety"; "deeper understanding of material in academic courses"; "more satisfying relationships with those close to you"; and "improved body posture."

Arousal reduction. Autonomic arousal was conceptualized as an abstract dimension, or factor, that reflects the common variance of many physiological response systems (Berlyne, 1967). Ideally, such a factor would be assessed by measuring a variety of different physiological indices, since any one index will share only some of its variance with the hypothetical construct of arousal. However, since we were not equipped to monitor several physiological channels, pulse rate alone was used as an approximate index of arousal. This particular variable

was selected because it has been shown to be affected by both muscle relaxation (Paul, 1969) and TM (Wallace, 1970), and it is generally considered to be one of the indices related to arousal level (Berlyne, 1967).

The extent of pulse rate reduction achieved by subjects while meditating, practicing muscle relaxation, or for the control subjects, simply resting with eyes closed, was monitored during an 18-min testing period. Subjects' pulses were measured by a light-sensitive plethysmograph (Grass Model 76604) attached to the index finger of the nondominant hand. The transduced signal was recorded by a Grass Model 76101 polygraph. The pulses were recorded for 20 sec immediately before the beginning of the testing period and for another 20 sec between 17 min 25 sec and 17 min 45 sec of the period.

Psychological maladjustment. Rotter's Incomplete Sentences Test, the measure of overall psychological maladjustment, was scored by the first author using Rotter and Rafferty's (1950) scoring manual; he was blind to both group and time of administration.

Trait anxiety (self-report). Each week of the experiment, subjects completed Zuckerman's (1960) Adjective Check List (ACL) scale of anxiety with instructions to check the adjectives that described how they felt "this week." In addition, the S-R Inventory of Anxiousness (Endler, Hunt, & Rosenstein, 1962) was administered on four occasions. This instrument requires the subject to indicate on a 5-point scale how strongly he or she would react with each of 14 common "anxiety" responses (sweaty palms, nausea, etc.) in each of 11 different situations. The ratings were summed to yield an overall index of trait anxiety.

Trait anxiety (behavioral). In response to the need for a behavioral measure to supplement the self-report measures of trait anxiety, the Behavioral Anxiety Measure (BAM) was developed by modifying a procedure used by Rehm and Marston (1968). The BAM consists of 10 tape-recorded situations, each of which contains a short description of a social situation followed by a line of dialogue spoken by someone in the situation. The subject's task is to respond to the line of dialogue. Thirty seconds were allowed for each situation and response unit. Subjects' performances were videotaped through a one-way mirror and rated using a checklist of anxiety indicators derived from that of Paul (1966).

The details of the procedure were as follows: The experimenter first presented written instructions, which emphasized that the subject was to respond "just as you would if the situation were actually happening" and that his or her performance would be rated for "overall psychological adjustment." The experimenter, who remained in the room with the subject throughout the testing, then played a demonstration tape of three situations and gave a prearranged sample reply for each.

Two sets of situations, each with a male and a female form, were prepared from an initial pool of 48 situations representing the four general areas of professor-student, parent-student, heterosexual, and same-sex peer relationships. This pool was submitted to 46 male and 42 female introductory psychology students, who rated each situation for how uncomfortable it would make them. The final sets of situations were selected so that the four forms (A-male, A-female, B-male, B-female) would be nearly equal in mean discomfort rating. A

sample situation (professor-student) is: "You have a large, 9:00 lecture class that you tend to be late for. Usually the professor just gives you a dirty look, but this morning as you enter the lecture hall she says, 'alright, how come you're late *again* today?'"

Nine relatively specific behaviors were chosen as anxiety indicators: head held downcast, abrupt head movements, swaying, extraneous arm and hand movements, arms held rigidly, hands restrained, failure to reply, blocking of speech, and extraneous comments. The raters practiced scoring videotapes of volunteer undergraduates until interrater reliability exceeding 70% was established for each of the nine items.

Three or four raters, blind to the subjects' group memberships, viewed each videotape and made two ratings for each of the 10 situation-response units. The first rating was based on the subject's behavior during the first 15 sec of the interval, and the second was based on the behavior of the final 15 sec. The score for the test was simply the sum of the number of 15-sec blocks in which each behavior was observed.

Procedure

Subjects received groups of measurements at six testing sessions: pretreatment, pretraining, posttraining, midtreatment, Posttreatment 1 and Posttreatment 2. (Posttreatment 1 and Posttreatment 2 were separated by approximately 1 week.) There was a delay of approximately 2 weeks between the pretreatment measures and the measures administered immediately before training (pretraining). The Locus of Control scale, the Social Desirability Scale, the first questionnaire, and the Incomplete Sentences Test were group administered at the pretreatment session; the Locus of Control scale and the Incomplete Sentences Test were readministered at Posttreatment 2, along with the second questionnaire. The S-R Inventory of Anxiousness was administered at pretraining, posttraining, midtreatment, and Posttreatment 1. The BAM was administered at pretraining and Posttreatment 1, and at Posttreatment 1 it was followed by the arousal-reduction measurement. All subjects completed a weekly ACL anxiety scale. In addition, subjects in the two treatment groups completed the expectancy questionnaire four times during their training and again at Posttreatment 2, and they provided a weekly report of the number of times they had meditated or relaxed.

The experimenters for the BAM were three male and three female undergraduates. A randomly selected half of the males and females within each group received BAM Set A at pretraining and Set B at Posttreatment 1 and the remaining subjects received the sets in the opposite order.

In accordance with the Student International Meditation Society's requirement of 15 days of abstinence from drugs prior to initiation, all subjects were asked to refrain from using drugs for the 2 weeks separating the pretreatment testing and the first day of training.

Results

The results of the study are presented in the following order: behavioral anxiety measure,

self-report anxiety measures, other outcome variables, and possible origins of differential treatment effects.

Behavioral Anxiety Measure

Interrater reliabilities for the scores on the BAM were computed using Pearson product-moment correlations. The mean reliabilities at pretraining and Posttreatment 1 were, respectively, .80 and .92. Using the Spearman-Brown formula for three independent raters, the reliabilities of the averaged scores were estimated to be .92 at pretraining and .97 at Posttreatment 1.

Test-retest reliabilities were computed using the control subjects' scores, and the reliability of the BAM scores was found to be satisfactory ($r = .58, p < .05$).

The BAM scores were analyzed in a repeated measurements analysis of variance design with one between (groups) and one within (time) factor.³ The time effect alone was significant, $F(1, 57) = 29.08, p < .001$. Inspection of the means revealed that there was an overall tendency to exhibit less anxiety at the second BAM testing but that there were no between-group differences in this regard. Thus, the behavioral anxiety measure provided no evidence of treatment effects of either muscle relaxation or TM on subjects' trait anxiety.

Self-report Anxiety Measures

The weekly ACLs were averaged over intervals of 2 weeks to obtain six biweekly average ACL scores for each subject; the first 2-week interval extended through the first day after the personal instruction of treated subjects and essentially constitutes a pretraining baseline.

A repeated measures analysis of variance disclosed a main effect for time, $F(5, 284) = 5.33, p < .001$, and a trend toward a groups effect that nearly reached the conventional significance level, $F(2, 57) = 3.06, p = .055$. Although the control group consistently reported greater anxiety than the other two groups,

³ All analyses were carried out using an unweighted means solution for unequal *N*s. Missing data were estimated by the Data-Text program, and error degrees of freedom were adjusted accordingly.

Table 1

Means of Total Scores on the S-R Inventory of Anxiousness by Group and Sex

Group	n	Administration			
		Pretraining	Posttraining	Midtreatment	Posttreatment 1
Muscle relaxation					
Males	10	336.0	331.0	325.7	332.4
Females	10	371.0	374.5	355.6	351.9
Transcendental meditation					
Males	10	353.6	328.3	304.6	294.2
Females	9	384.2	369.3	346.1	344.3
Control					
Males	10	362.4	348.9	334.8	337.1
Females	11	434.4	435.9	438.4	436.6

there was a downward trend in anxiety across the groups. A trend analysis (Meyers, 1972) indicated that there was a significant linear component to this trend, $F(1, 57) = 10.41$, $p < .001$.

The total scores for the S-R Inventory of Anxiousness were analyzed in a repeated measures analysis of variance design with groups and sex as between-subjects factors. Significant effects were found for sex, $F(1, 54) = 12.20$, $p < .001$; groups, $F(2, 54) = 4.13$, $p < .05$; time, $F(3, 161) = 14.56$, $p < .001$; and the Groups \times Time interaction, $F(6, 161) = 3.53$, $p < .01$. It can be seen from the means (Table 1) that females consistently reported greater anxiety than males and that the control group consistently scored higher than the two treatment groups. The most interesting finding was that although the muscle relaxation and control groups' scores remained essentially unchanged over time, the TM group reported substantial decreases in anxiety. The absence of a triple interaction involving sex indicates that the differential treatment effect found in these data was present equally for males and females. Therefore, the sex factor was dropped from a trend analysis performed to investigate the nature of the significant Groups \times Time effect. The S-R inventory was assumed to have been administered to all subjects on the 1st, 11th, 37th, and 71st days of the experiment, and orthogonal linear ($-29, -19, 7, 41$) and quadratic ($3.85, -1.00, -6.16, 3.31$) coefficients were derived (Meyers, 1972). The linear component of the interaction was found to be significant, $F(2, 57) = 4.72$, $p < .02$, but the quadratic component was not, which indicates

that the decreases in the TM group's reported anxiety occurred steadily over the 9-week treatment period.

Thus, of the two self-report measures of trait anxiety, the S-R inventory provided evidence of a treatment effect of TM, but the ACL did not.

Other Outcome Variables

Maladjustment scores from the Incomplete Sentences Test administered at pretreatment and Posttreatment 2 were analyzed in a repeated measures analysis of variance with groups as a between-subjects factor. The only significant effect was time, $F(1, 56) = 8.18$, $p = .006$, which was associated with an overall decrease in maladjustment scores.

Analysis of subjects' locus of control scores disclosed no significant effects, indicating that the groups neither differed initially nor changed significantly over the course of the experiment.

Subjects' stated frequencies of drunkenness and marijuana use "in the recent past" were collected at pretreatment and Posttreatment 2 and were subjected to another repeated measures analysis of variance. No significant effects were found for either frequency of drunkenness or frequency of marijuana use.

At Posttreatment 2, subjects in the two treatment groups were asked to rate on a 5-point scale how much benefit they felt they had gained in each of eight areas: academic performance, interpersonal relations, health, condition of body and nervous system, decreased drug use, increased energy and vitality, overall happiness, decreased anxiety, and self-

esteem. The first, third, and fifth points on the scale were labeled, respectively, *not at all*, *a moderate amount*, and *a great deal*. The mean ratings of the muscle relaxation and TM groups were compared by *t* tests, and it was found that the only significant difference was for academic performance, with TM subjects reporting greater benefits, $t(18) = 2.51$, $p < .05$. It is noteworthy that the TM group tended to report only "moderate" benefits, with mean ratings ranging from 2.20 (decreased drug use) to 3.32 (increased energy and vitality).

Possible Origins of Differential Treatment Effects

Additional analyses were performed to determine if there were unintended differences between the treatment groups that might have been responsible for the observed differential treatment effect.

Unfortunately, the posttraining arousal reduction data had to be discarded because of equipment problems that remained undetected for numerous subjects. The analysis of variance of the Posttreatment 1 data revealed an overall decrease in pulse rate (M change = 2.8 beats per minute) from the beginning to the end of the 18-minute testing period, $F(1, 57) = 13.78$, $p < .001$, but there was no evidence that meditation produced greater reduction in arousal than muscle relaxation or simply sitting with eyes closed.

The treated subjects' weekly frequencies of practice were averaged over 2-week intervals and analyzed in a repeated measures analysis of variance. Main effects for groups, $F(1, 37) = 6.70$, $p < .05$, and time $F(4, 147)$, $p < .001$ were found, but their interaction was not significant. Both groups decreased in frequency after the first 2 weeks, and although the TM group maintained a consistently higher frequency, the magnitude of the average difference was not great (for TM, $M = 11.5$; for muscle relaxation, $M = 10.2$).

Subjects' total expectancy scores after their training (posttraining) and at the end of the experiment (Posttreatment 2) were used as an index of the nonspecific effects of treatments. Repeated measures analysis of variance disclosed a significant effect for time, $F(1, 37) = 13.52$, $p < .001$, and a trend toward a groups effect, $F(1, 37) = 2.89$, $p = .10$. Both

groups' expectancies decreased over the treatment period. The trend in the groups factor reflected the TM group's higher expectancies and suggests that the treatments may not have been equal in their nonspecific effects.

To determine if the variables on which differences were found between the treatment groups were actually associated with the observed treatment effect of TM, correlations were computed between these variables and residual change scores on the S-R inventory. The change scores were computed from the TM subjects' pretraining and Posttreatment 1 S-R inventory scores, using the linear regression coefficient derived from the entire sample. The correlations between the treatment effect, as measured by the residual change scores, and variables we thought might be related to treatment effect were average frequency of practice, $-.23$; total expectancy, $-.26$; social desirability, $-.20$; psychological adjustment, $.06$; and locus of control, $.53$ ($ps < .05$). It is apparent that the two variables on which the treatments differed, frequency of practice and expectancy, were not significantly related to TM subjects' decreases in trait anxiety.

Discussion

Interpretation of these findings is made difficult by the inconsistency of the results obtained with the three measures of trait anxiety. In particular, it is unclear whether the S-R inventory measured some change in TM subjects to which the other instruments were insensitive or whether that apparent change was an artifact of measurement.

One possibility is that the TM treatment effect was a result of subjects' perceptions that this was "an experiment on TM" and their responses to the implicit demands (Orne, 1969) to report decreases in anxiety. On the other hand, the fact that Social Desirability scores did not predict residual gain scores within the TM groups argues against this hypothesis, since it would be expected that subjects more highly motivated to win the experimenter's approval would report larger decrements. A second possibility is that the decline in anxiety scores indicated changes in subjects' self-images rather than actual changes in trait anxiety. It is difficult to reconcile the self-image interpre-

tation of the S-R inventory results with the failure to find a treatment effect on the anxiety ACL, however, since the ACL is a classic format for eliciting self-images and should have been sensitive to beliefs such as "I don't get so tense anymore." Furthermore, there is some evidence (Otis, 1974) that TM does not affect subjects' self-images.

Thus, although not conclusive, the available evidence favors the view that the S-R inventory measured a real decrease in meditators' trait anxiety. How, then, can we reconcile the positive results obtained from this instrument with the negative results obtained with the ACL and the BAM? A cue is provided by the fact that analyses performed using the response factor scores (Endler et al., 1962) in place of the overall score on the S-R inventory found TM to be more effective than the other conditions in reducing scores on the "distress" factor but not on the "autonomic" factor. Recalling Lang's (1969) distinction among the self-report, behavioral, and autonomic dimensions of anxiety, it seems reasonable to suppose that the distress factor corresponds to the self-report component of anxiety and that our S-R inventory results primarily reflect changes in subjective distress. This implies that although TM may have had no effect on the behavioral or autonomic components of anxiety, or on global self-images of anxiety proneness, it may have reduced subjectively experienced distress in the specific situations tapped by the S-R inventory. If this interpretation were valid, it would follow that TM first affects subjective distress, and, at least over short treatment periods, does not alter overt behavior.

Even if one accepts the conclusion that TM was effective in reducing some aspect of anxiety, it can still be argued that the greater reduction in anxiety reported by TM subjects was due to the operation of a stronger placebo effect rather than to a *specific* effect on TM, since training in TM produced higher expectancies for benefits. The plausibility of this interpretation of the S-R inventory results is reduced by three findings: First, though a significant relationship between expectancy and outcome measures would be expected if the treatment were functioning primarily as a placebo, TM subjects' expectancies immediately after the training period did not signifi-

cantly predict residual change scores on the S-R inventory. Second, subjects' expectancies decreased over the treatment period in which their reported trait anxiety decreased, and it seems unlikely that a treatment could have a strong placebo effect while subjects' confidence in the treatment was waning. Finally, the decreases in anxiety were not abrupt, as would be expected from a placebo effect, but were gradual and continuous, as would be expected from a true treatment effect.

This moderately supportive evidence for the existence of a specific treatment effect of TM is not consistent with Smith's (1976) study, which found TM to be no more effective than a nonspecific effects control treatment, which consisted of sitting passively in a chair. This inconsistency could reflect differences in the measures of anxiety used in the two studies, differences in responses to treatment as a function of the populations studied, or differences between the nonspecific effects control treatments. Smith's subjects volunteered specifically to receive treatment for anxiety, whereas the majority of subjects in the present study were hoping to receive training in TM. It is possible that the groups who received what they were seeking—Smith's two treatment groups and our TM group—responded with reports of decreased anxiety. Another possibility is that the high dropout rate in Smith's study was not random but instead reflected a Treatment \times Subjects interaction in which different subpopulations responded specifically to each treatment, and the nonresponders in each group tended to drop out. Finally, Smith's control treatment may have incorporated a therapeutic ingredient that was absent in this study's muscle relaxation treatment, for example, the pairing of relaxation with continued cognitive activity, the feeling that the technique was easy and enjoyable to practice, or a placebo effect fully as strong as that of TM.

In contrast to previous studies, the present study found TM to have no effect on subjects' scores for locus of control, psychological maladjustment, or frequencies of drunkenness and marijuana use. However, with the exception of the studies by Smith (1975) and Shafii et al. (1974), the earlier results are compromised by weak designs and measures of questionable validity. The results of the last two studies are

not directly comparable to those reported here, because they were obtained with different populations, different measures, and over longer treatment periods. Until further research clarifies the situation, the most that can be said with assurance is that there is at present as much negative as positive evidence concerning the effects of TM on maladjustment and drug use, and there is no sound evidence that it affects locus of control.

In summary, though the results of this study provide some support for the hypothesis that TM is specifically effective in reducing normal college students' experiences of anxiety, it must be remembered that negative results were obtained with the behavioral measure of trait anxiety and with measures of locus of control, psychological maladjustment, and frequency of drunkenness and marijuana use, and that the technique tended to be rated as only "moderately" helpful with general life problems. It appears, therefore, that TM has been oversold by its proponents, and unless it is shown that long-term practice does lead to great benefits, it should be considered irresponsible to advertise TM as a panacea.

References

- Ballou, D. Transcendental meditation research: Minnesota State Prison. In D. P. Kanellakos & P. C. Ferguson (Eds.), *The psychobiology of transcendental meditation*. Los Angeles: Maharishi International University Press, 1973. (Abstract)
- Berlyne, D. Arousal and reinforcement. In D. Levine (Ed.), *Nebraska Symposium on Motivation*, 1967. Lincoln: University of Nebraska Press, 1967.
- Crowne, D. P., & Marlowe, D. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 1960, 24, 349-354.
- Endler, N. S., Hunt, J. McV., & Rosenstein, A. J. An S-R inventory of anxiousness. *Psychological Monographs*, 1962, 76(17, Whole No. 536).
- Ferguson, P. C., & Gowan, J. The influence of the practice of transcendental meditation on anxiety, depression, aggression, neuroticism, and self-actualization. In D. P. Kanellakos & P. C. Ferguson (Eds.), *The psychobiology of transcendental meditation*. Los Angeles: Maharishi International University Press, 1973. (Abstract).
- Hjelle, L. A. Transcendental meditation and psychological health. *Perceptual and Motor Skills*, 1974, 39, 623-628.
- Jacobson, E. *Progressive relaxation*. Chicago: University of Chicago Press, 1938.
- Lang, P. The mechanics of desensitization and the laboratory study of human fear. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- Meyers, J. L. *Fundamentals of experimental design*. Boston: Allyn & Bacon, 1972.
- Orme-Johnson, D. W., Arthur, G., Franklin, L., O'Connell, J., & Zold, T. Transcendental meditation and drug abuse counselors. In D. P. Kanellakos & P. C. Ferguson (Eds.), *The psychobiology of transcendental meditation*. Los Angeles: Maharishi International University Press, 1973. (Abstract)
- Orne, M. T. Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press, 1969.
- Otis, L. S. If well-integrated but anxious, try TM. *Psychology Today*, 1974, 7, 45-47.
- Paul, G. L. *Insight vs. desensitization in psychotherapy*. Stanford: Stanford University Press, 1966.
- Paul, G. L. Physiological effects of relaxation training. *Journal of Abnormal Psychology*, 1969, 74, 425-437.
- Rehm, L. P., & Marston, A. R. Reduction of social anxiety through modification of self-reinforcement: An instigation therapy technique. *Journal of Consulting and Clinical Psychology*, 1968, 32, 565-574.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Rotter, J. B., & Rafferty, J. E. *Manual for the Rotter Incomplete Sentences Blank College Form*. New York: Psychological Corporation, 1950.
- Seeman, W., Nidich, S., & Banta, T. Influence of transcendental meditation on a measure of self-actualization. *Journal of Counseling Psychology*, 1972, 19, 184-187.
- Shafii, M., Lavelly, R., & Jaffee, R. Meditation and marijuana. *American Journal of Psychiatry*, 1974, 131, 60-63.
- Smith, J. C. Meditation as psychotherapy: A review of the literature. *Psychological Bulletin*, 1975, 82, 558-564.
- Smith, J. C. Psychotherapeutic effects of transcendental meditation with controls for expectation of relief and daily sitting. *Journal of Consulting and Clinical Psychology*, 1976, 44, 630-637.
- Spielberger, C. D. Anxiety as an emotional state. In C. D. Spielberger (Ed.), *Anxiety*. New York: Academic Press, 1972.
- Wallace, R. K. Physiological effects of transcendental meditation. *Science*, 1970, 167, 1751-1754.
- Zuckerman, M. The development of an affect adjective check list for the measurement of anxiety. *Journal of Consulting Psychology*, 1960, 24, 456-462.

Received April 25, 1977 ■

Personality Correlates of Continuation and Outcome in Meditation and Erect Sitting Control Treatments

Jonathan C. Smith
Roosevelt University

In a 6-month double-blind study, 49 anxious college student volunteers were assigned to transcendental meditation (TM) and 51 to a control treatment, periodic somatic inactivity (PSI). The control treatment was carefully designed to match the form, complexity, and expectation-fostering aspects of TM, but it incorporated an exercise that involved sitting erect with eyes closed twice daily rather than sitting and meditating. For each treatment 30 demographic and pretest personality variables were correlated with continuation in treatment and outcome defined in terms of trait anxiety change scores. As predicted, the TM dropout was more disturbed and less self-critical than the person who continued meditating. For TM, outcome correlated significantly with anxiety, Sizothymia (16 Personality Factor Questionnaire, Factor A), and Autia (16 Personality Factor Questionnaire, Factor M). Contrary to what was predicted, there was virtually no overlap between the variables correlated with continuation and outcome for TM and for PSI. It is concluded that differing treatment rationales rendered the treatments appealing, credible, and effective for different types of individuals.

Meditation research seems destined to repeat the sins of a generation of psychotherapy outcome research. In both areas the tendency has been to ask "does it work?" As Bergin (1971) has chronicled, this amorphous question has littered the journals with controversial and ambiguous results. Instead, Paul (1967) suggested:

The question towards which all outcome research should ultimately be directed is the following: *What treatment, by whom, is most effective for this individual with that specific problem, and under which set of circumstances?* (p. 111)

Meditation researchers have tended to ignore such questions of specificity (Smith, Note 1).

A surplus of studies show that individuals generally display reductions in trait anxiety after learning meditation (Smith, 1975b). However, not everyone benefits, and up to half discontinue (Smith, 1976; Otis, Note 2). To date no one has systematically explored

the type of person who gains from meditation, and only one person, Otis (Note 2), has looked at the meditation dropout. On the basis of two questionnaire studies, Otis found that subjects who stop practicing transcendental meditation (TM) compared with those who continue feel "less positive about themselves," have "more serious problems," and are more "withdrawn, irritable, and anxiety ridden." In addition, he suggested that subjects who continue with TM are somewhat less disturbed, although they may admit to more problems; that is, they may be more self-critical.

The present study examined characteristics of subjects who drop out of TM and subjects who continue practicing for up to 6 months and who display significant reductions in trait anxiety. On the basis of Otis's findings, it was predicted that subjects who continue practicing TM, compared with those who discontinue, are at pretest less disturbed, less anxious, and less withdrawn but more self-critical. Also, a previous study based on the same subject sample used in the present study (Smith, 1976) concluded that TM and a yogalike treatment involving sitting erect are equally

Requests for reprints should be sent to Jonathan C. Smith, Department of Psychology, Roosevelt University, 430 South Michigan Avenue, Chicago, Illinois 60605.

psychotherapeutic, and that the therapeutic processes operating in both are the same—some combination of expectation of relief and daily sitting. For this reason it was predicted that the correlates of continuation and outcome are the same for TM and the treatment involving sitting erect.

Method

Procedure

The present study used data collected in a previous study (Smith, 1976) that compared the psychotherapeutic effects of TM and a control treatment, periodic somatic inactivity (PSI). Subjects were 100 (51 male, 49 female) Michigan State University students who were suffering from a high level of trait anxiety. All were carefully screened for motivation, were not involved in psychotherapy, and had at no time practiced meditation or yoga. Mean age was 22.

All subjects were pretested, and 49 were randomly assigned to TM and 51 to PSI. TM was taught by two official TM instructors from the Students' International Meditation Society and was identical to ordinary TM in every respect except that it was offered free. The TM technique involves sitting erect with feet flat on the floor and eyes closed while passively and continuously attending to a special thought called a "mantra." This is done for 15–20 minutes twice daily. Complete TM instruction includes two introductory lectures that outline supporting theory and research, a 15-day drug fast, standardized individual instruction, 3 days of follow-up instruction and discussion, and monthly follow-up checking.

PSI was designed to control for the potentially therapeutic effects of daily sitting and expectation of relief. Specifically, the treatment matched TM in every respect with one exception—instead of sitting and meditating, the instructions were to simply sit. Subjects were told that while in this position they could think about anything (even worry) and the technique would still work. Like TM, PSI instruction began with two introductory lectures that outlined a rationale explaining why sitting twice daily should be an immensely effective cure for most forms of psychopathology. In addition, bogus research was presented to support the claims made. Between lectures, subjects participated in a 15-day fast from illegal drugs. After the fast and lectures, subjects were individually initiated and met for 3 days of follow-up checking. PSI was taught double blind; both the subjects and the instructor were deceived into believing that the treatment was legitimate and widely researched and not a bogus control treatment.

One important feature of PSI was its rationale. Care was taken to construct a rationale that was credible and complex. To enhance credibility, actual psychological concepts and research were woven together in a superficially elegant manner. That not one component of PSI theory was false or deceptive (although supporting "process" and "outcome" research was faked) makes

it unique among bogus treatment rationales. A summary of the rationale given to subjects is presented below (Smith, 1975a):

Built into life are factors that disrupt inner calm and generate and maintain anxiety. Research has shown that one of these factors is the desynchronization of circadian rhythms, daily rhythmic changes in physiological functioning. PSI works to bring circadian rhythms into synchrony.

The way PSI works is complex. All physical activity, no matter how small, generates a fatiguelike and stresslike nonspecific physiological by-product called *reactive inhibition*. Simple physical inactivity tends to trigger the automatic dissipation of reactive inhibition. Such dissipation appears physiologically as a decrease in physiological activity and as a small dip or signature in the constellation of circadian rhythms. PSI involves remaining physically inactive for 15 to 20 minutes at the same time each day. The result is that regular inactivity-induced signatures appear at and become classically conditioned to the same point in one's circadian rhythms each day. As one continues practicing PSI, conditioning continues, overlearning occurs, "dips become conditioned onto dips," and gradually, and automatically, the associated physiological changes become deeper and deeper.

The regular appearance of inactivity-induced signatures in circadian rhythms serve as zeitgeber, stimuli that pull and keep circadian rhythms in synchrony. PSI thereby functions to pull and keep circadian rhythms in synchrony, and as a result reduces anxiety and increases psychological well-being.

Periodic inactivity is the single commonality among a variety of highly effective growth and therapy techniques including progressive relaxation, biofeedback training, autogenic therapy, self-hypnosis, meditation, and yoga. However, since PSI incorporates only the essentials of these techniques, and does away with all the unnecessary and cumbersome extras associated with them, it is in fact more effective and efficient.

Both treatments continued for 6 months, after which subjects were posttested and debriefed. Twenty TM and 24 PSI subjects reported for posttesting. To this sample were added 2 TM and 3 PSI subjects who were not available for 6-month posttesting but who did take an abbreviated posttest after 3½ months. These subjects were included to increase sample size, since there is strong evidence that 6 months of TM is no more therapeutic than 3½ months (Smith, 1975a).

Outcome and Continuation Measures

Trait anxiety was selected as the main outcome variable, specifically pretest-posttest difference scores on the Anxiety Trait (A-Trait) scale of the State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1970). Anxiety was selected because it is the most widely studied trait in meditation research (Smith,

1975b). Although proponents of TM claim that their technique has a desirable impact on other variables, notably self-esteem, psychosis, self-actualization, creativity, and even intelligence (Glueck & Stroebel, 1975; Kanellakos & Ferguson, 1973; Orme-Johnson, Domash, & Farrow, 1974), the supporting evidence is scanty. Indeed, TM subjects in the present study displayed no change on any of these variables (Smith, 1975a).

In addition, at posttest subjects were asked to estimate how often they had practiced each month throughout the project. Subjects were considered to have continued practicing if they practiced at least once during the last month.

Demographic and Pretest Personality Measures

Before the onset of the project, each subject was given the STAI A-Trait scale, the Epstein-Fenz Manifest Anxiety Scale (Fenz & Epstein, 1965), the 16 Personality Factor Questionnaire Forms A and B (16 PF; Cattell, Eber & Tatsuoka, 1970), the IPAT Neuroticism Scale Questionnaire (Scheier & Cattell, 1961), the Tennessee Self-Concept Scale (Fitts, 1965), and the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1964). The Epstein-Fenz test was scored for symptoms of autonomic arousal and symptoms of striated muscle tension. The three Cattell tests were pooled to increase factor reliability and were scored for the 16 primary source traits. On the Tennessee scale, the Total Positive, Psychosis, Personality Disorder, Personality Integration, Defensive Positive, and Self-criticism scales were used. In addition subjects indicated their sex and age and stated if they had at any time prior to the project considered psychotherapy or meditation and yoga. In sum, outcome and continuation were correlated with 30 variables.

Results

Tables 1 and 2 show the correlations of demographic and pretest personality variables with outcome and continuation for TM and PSI subjects. For TM subjects, outcome correlated significantly ($r \geq .396$, $p \leq .047$) with (ranked in order of significance): not having considered psychotherapy prior to the onset of the project, Sizothymia (16 PF Factor A), Autia (Factor M), anxiety (STAI A-Trait), Weaker Superego Strength (Factor G), and lack of Personality Integration.

Continuation with TM correlated significantly ($r \geq .397$, $p \leq .034$) with a low degree of Psychoticism and a high degree of Self-criticism, as well as with having considered psychotherapy prior to the onset of the project.

For PSI the results were quite different. Outcome correlated significantly ($r \geq .462$, $p \leq .027$) with Shrewdness (Factor N). Con-

tinuation correlated significantly ($r \geq .338$, $p \leq .042$) with Alaxia (Factor L), Shrewdness (Factor N), Desurgency (Factor F), Untroubled Adequacy (Factor O), High Strength of Self-sentiment (Factor Q_s), and Ego Strength (Factor C).

Discussion

One must be extremely cautious when drawing conclusions from research involving many variables and few subjects. For this reason I chose to give particular credence only to those variables that correlated most highly with outcome and continuation and have in previous research displayed greatest validity, reliability, and immunity to the effects of motivational distortion.

An intriguing picture emerges of those individuals who continue with TM and display the greatest reduction in trait anxiety. Not only are they anxious, but they score high on 16 PF Factors Sizothymia and Autia. Cattell and his colleagues (Cattell, 1957; Cattell et al., 1970) describe Sizothmic individuals as "reserved, detached, critical, cool, aloof," and "stiff." Emotionally, they are "flat" or "cautious." They tend to be critical, precise, and skeptical, and like working alone with things or words rather than with people. In interpreting this factor, Cattell (1957) hypothesizes that it reflects a "steadiness in purpose and a high level of interest in symbolic and subjective activity...a secondary result of blocking of easy interaction with the changing external world" (p. 180). In light of the apparent introversion of those who benefit from TM, it is not surprising that they tend not to have considered psychotherapy as a treatment option.

Those who benefit from TM also score high on Autia. These individuals tend to be unconventional and interested in "art, theory, basic beliefs" and "spiritual matters." However, their most important characteristic is what Cattell variously describes as a tendency to be "imaginatively enthralled by inner creations," "charmed by works of the imagination," and "completely absorbed" in the momentum of their own thoughts, following them "wherever they lead, for their intrinsic attractiveness and with neglect of realistic

considerations." Cattell speculated that fundamental to Autia may be a capacity to disassociate and engage in "autonomous, self-absorbed relaxation."

A quite different picture emerges of those who continue with and benefit from PSI. They tend to score high on 16 PF Factor N. Such individuals, according to Cattell, tend

Table 1

Demographic and Pretest Personality Correlates of A-Trait Change Scores for Transcendental Meditation (TM) and Periodic Somatic Inactivity (PSI)

Variable	TM			PSI		
	r	n	p	r	n	p
Sex ^a	.297	19	.109	.021	18	.467
Age	-.120	18	.317	.206	18	.206
Considered meditation ^b	.070	19	.388	.366	18	.067
Considered therapy ^c	-.546	19	.008	.387	18	.056
A-Trait	.488	19	.017	.359	18	.072
SSMT	-.136	19	.289	.086	18	.367
SAA	.021	19	.466	.106	18	.338
Factor A ^d	-.543	18	.010	.272	18	.138
Factor B	-.064	18	.400	.132	18	.301
Factor C	-.133	18	.300	-.013	18	.480
Factor E	-.305	18	.110	.104	18	.341
Factor F	-.395	18	.052	-.037	18	.443
Factor G	-.485	18	.021	.005	18	.492
Factor H	-.309	18	.106	.181	18	.236
Factor I	.378	18	.061	.296	18	.117
Factor L	.044	18	.431	.150	18	.276
Factor M	.519	18	.014	-.001	18	.499
Factor N	.286	18	.125	.462	18	.027
Factor O	.183	18	.234	-.197	18	.216
Factor Q ₁	-.107	18	.336	-.345	18	.081
Factor Q ₂	.288	18	.123	-.066	18	.397
Factor Q ₃	-.001	18	.498	.009	18	.486
Factor Q ₄	.148	18	.279	-.220	18	.190
Marlowe-Crowne	.208	19	.197	.262	18	.147
Defensive positive ^e	-.002	19	.497	.157	18	.267
Psychosis	.077	19	.378	.376	18	.062
Personality disorder	-.044	19	.429	.175	18	.243
Personality integration	-.396	19	.047	.117	18	.322
Self-criticism	-.361	19	.064	.034	18	.447
Total positive	-.044	19	.428	.187	18	.229

Note. A-Trait = Trait Anxiety scale of the State-Trait Anxiety Inventory; SSMT = symptoms of striated muscle tension; SAA = symptoms of autonomic arousal; Factor A = Sizothymia vs. Affectothymia; Factor B = Low Intelligence vs. High Intelligence; Factor C = Lower Ego Strength vs. Higher Ego Strength; Factor E = Submissiveness vs. Dominance; Factor F = Desurgency vs. Surgency; Factor G = Weaker Superego Strength vs. Stronger Superego Strength; Factor H = Thrextia vs. Parmia; Factor I = Harria vs. Premsia; Factor L = Alaxia vs. Protension; Factor M = Praxernia vs. Autia; Factor N = Artlessness vs. Shrewdness; Factor O = Untroubled Adequacy vs. Guilt Proneness; Factor Q₁ = Conservatism of Temperament vs. Radicalism; Factor Q₂ = Group Adherence vs. Self-sufficiency; Factor Q₃ = Low Self-sentiment Integration vs. High Strength of Self-sentiment; Factor Q₄ = Low Ergic Tension vs. High Ergic Tension.

^a Keyed so that 1 = male, 2 = female.

^b Keyed so that 1 = did not consider meditation prior to the project, 2 = did consider meditation prior to the project.

^c Keyed so that 1 = did not consider therapy, 2 = did consider therapy.

^d 16 Personality Factor Questionnaire (16 PF) factors were obtained by pooling scores from Forms A and B of the 16 PF with scores from the Neuroticism Scale Questionnaire.

^e This and the following variables were taken from the Tennessee Self-Concept Scale.

to have "exact calculating" minds and tend to be emotionally detached and disciplined, ambitious, and esthetically fastidious. They tend not to be gregarious or to get "warmly emotionally involved" with others.

My hypothesis that the TM dropout is disturbed, anxious, withdrawn, and somewhat lacking in self-criticism appears to be supported. Although only three dropouts reported

for posttesting, they had extremely high Psychoticism scores. Fitts (1965) reports the "normal limits" on Psychoticism to be from 34 to 54 and the average score of hospitalized psychotics to be 62. The average Psychoticism score of the TM dropout was 63 ($SD = 4.58$), whereas the average score of those who continued was 48.74 ($SD = 6.46$). In addition, the dropouts scored lower in Self-criticism.

Table 2

Demographic and Pretest Personality Correlates of Continuation for Transcendental Meditation (TM) and Periodic Somatic Inactivity (PSI)*

Variable	TM			PSI		
	<i>r</i>	<i>n</i>	<i>p</i>	<i>r</i>	<i>n</i>	<i>p</i>
Sex	.025	22	.456	-.158	27	.215
Age	.164	21	.239	.281	27	.078
Considered meditation	.208	22	.176	-.267	27	.090
Considered therapy	.397	22	.034	.158	27	.215
A-Trait	.180	22	.212	-.032	27	.436
SSMT	.021	22	.463	-.069	27	.367
SAA	.014	22	.476	-.039	27	.423
Factor A	.194	21	.200	-.310	27	.058
Factor B	-.170	21	.230	-.159	27	.214
Factor C	.012	21	.480	.338	27	.042
Factor E	-.094	21	.343	.003	27	.495
Factor F	.236	21	.152	-.440	27	.011
Factor G	-.028	21	.451	-.052	27	.398
Factor H	.056	21	.404	-.119	27	.278
Factor I	.211	21	.179	-.109	27	.294
Factor L	-.337	21	.067	-.548	27	.002
Factor M	.161	21	.243	.130	27	.260
Factor N	.038	21	.436	.455	27	.009
Factor O	.123	21	.297	-.377	27	.026
Factor Q ₁	-.232	21	.156	.156	27	.219
Factor Q ₂	-.329	21	.073	.267	27	.089
Factor Q ₃	-.108	21	.320	.342	27	.040
Factor Q ₄	.280	21	.109	-.298	27	.066
Marlowe-Crowne	-.034	22	.441	.093	27	.322
Defensive positive	-.105	22	.320	-.012	26	.476
Psychosis	-.586	22	.002	.039	26	.426
Personality disorder	-.054	22	.405	.135	26	.255
Personality integration	-.032	22	.443	-.044	26	.416
Self-criticism	.409	22	.029	-.193	26	.173
Total positive	.028	22	.450	.076	26	.357

Note. A-Trait = Trait Anxiety scale of the State-Trait Anxiety Inventory; SSMT = symptoms of striated muscle tension; SAA = symptoms of autonomic arousal; Factor A = Sizothymia vs. Affectothymia; Factor B = Low Intelligence vs. High Intelligence; Factor C = Lower Ego Strength vs. Higher Ego Strength; Factor E = Submissiveness vs. Dominance; Factor F = Desurgency vs. Surgency; Factor G = Weaker Superego Strength vs. Stronger Superego Strength; Factor H = Threictia vs. Parmia; Factor I = Harria vs. Premsia; Factor L = Alaxia vs. Protension; Factor M = Praxernia vs. Autia; Factor N = Artlessness vs. Shrewdness; Factor O = Untroubled Adequacy vs. Guilt Proneness; Factor Q₁ = Conservatism of Temperament vs. Radicalism; Factor Q₂ = Group Adherence vs. Self-sufficiency; Factor Q₃ = Low Self-sentiment Integration vs. High Strength of Self-sentiment; Factor Q₄ = Low Ergic Tension vs. High Ergic Tension.

* Keyed so that 1 = did not practice at least once during the last month of the project, 2 = did practice at least once during the last month of the project.

Again, a somewhat different picture emerges of those who stop PSI. Consistent with what was predicted, they score higher on 16 PF factors related to anxiety (Factors L, O, Q₃, and C). However, they do not score higher on Psychoticism or lower on Self-criticism. Tentatively, they appear to be suspecting and are prone to dwell on frustrations (Factor L), as well as being a bit naive and gregarious (Factors N and F).

Examining the overall pattern of these results, it appears that there is virtually *no* overlap between the variables correlated with continuation and outcome for TM and for PSI. Yet I had hypothesized that if the same therapeutic processes were operating in the two treatments, they should work for the same types of individuals. I propose that a key to interpreting this inconsistency can be found in the possible interaction between treatment rationale and treatment outcome.

The rationale given for a treatment is fast emerging as a variable that can mediate attention placebo and possibly actual treatment effects (Borkovec & Nau, 1972; McReynolds, Barnes, Brooks, & Rehagen, 1973; Rosen, 1975, 1976). A treatment with a rationale that lacks credibility can be less effective than a treatment with a credible rationale. A rationale that imparts the expectation that a set of procedures is not therapeutic, or does not constitute a treatment, can reduce the effectiveness of those procedures. In addition, different rationales might render treatments appealing, credible, and as a result, effective for different types of individuals. For example, Fish (1973) described a case in which a patient from the hippie counterculture responded to systematic desensitization when it was presented as a consciousness-raising technique, whereas an engineer responded to the same technique when it was explained in terms of reciprocal inhibition.

Following what is common procedure in outcome research using attention placebo controls, highly credible but *different* rationales were given for TM and PSI. These rationales may have rendered the treatments credible, appealing, and effective for different types of individuals. The TM rationale, a religious-philosophical system from Hindu Vedantic tradition, can be summarized briefly:¹

The TM technique involves passively and continuously attending to a meaningless thought or word called a mantra. As this thought repeats in meditation, it is experienced at progressively earlier phases in its development as increasingly subtle, fine, and charming. This process is similar to following bubbles arising in an ocean to their source in the ocean bed.

As one attends to the mantra, distracting thoughts, images, and feelings spontaneously emerge. These are treated with detached acceptance; one simply favors the mantra as soon as he or she recognizes that attention has wandered. Such distractions are normal and indicate that stress is being released.

As a thought is experienced at developmentally earlier stages, relaxation increases, which, in turn, dissolves neurological knots of tension. One's mind becomes increasingly still, like a rippleless pond or the depths of an ocean. One approaches the source of thought, the eternal and absolute field of transcendental being, creative intelligence. All aspects of life are thereby enriched and improved.

It takes little imagination to see how the TM rationale may be highly appealing and credible to Sizothymic and Autic individuals. As described earlier, such people tend to be emotionally cool, steady, and detached, and might find the TM metaphors relating to stillness and detachment appealing. They tend to be "charmed" by "inner creations," a characteristic that aptly summarizes the gist

¹ The specific content of the rationale given in TM lectures and instruction is not available to the non-TM researcher. However, after interviewing two TM instructors and 15 practitioners, I concluded that much of the rationale is similar to accounts published in the popular press by TM instructors and advocates Maharishi Mahesh Yogi (1968) and Bloomfield, Cain, and Jaffe (1975). The rationale presented in this article is derived from the accounts of Maharishi and Bloomfield et al.

It should be noted that the TM program is shrouded with considerable secrecy. Not only are the specific content of lectures and instruction unavailable to the outside researcher, but TM practitioners are urged not to describe their technique or training to others. If psychologists wish to learn TM, they are asked to promise not to divulge the technique. Secrecy tends to be reinforced by the requirement that TM research proposals must first be approved by the TM organization. In my opinion, these restrictions pose serious ethical problems for the TM researcher. As exemplified in the present study, such secrecy limits the extent to which procedural components of the TM program can be isolated and investigated. At least one study on TM has been prematurely terminated partly because of the ethical difficulties associated with secrecy (White, 1976).

of the claimed TM process. And they tend to be unconventional and interested in art, theory, basic beliefs, and spiritual matters. The TM philosophy is blatantly spiritual and relies heavily on visual "artistic" metaphor.

In contrast to the TM rationale, the PSI rationale was highly intricate and mechanistic. The person who benefits from PSI was described as "shrewd, astute, exacting, calculating, ambitious," and having complex tastes. Such a person might well find the scientific-sounding precision and ambitious complexity of the PSI rationale highly credible and appealing.

Another interpretation of these findings is that the treatment processes basic to TM and PSI are in fact different and are effective for different types of individuals. Indeed, the act of meditation has frequently been claimed to have therapeutic properties (Smith, 1975b), and Pratap (1972) argues that sitting erect with eyes closed may in itself be therapeutic. However, if sitting and meditation are equally therapeutic for different individuals, then one would expect more people to benefit from TM than from PSI, since TM is done in a seated position. This was clearly not the case. The most direct way out of this inconsistency is to make the somewhat awkward assumption that something in TM, perhaps some undisclosed aspect of the TM rationale, suppresses the therapeutic impact of erect sitting.

However they may be interpreted, the findings of this experiment clearly show that personality characteristics are correlated with continuation and outcome in meditation. I propose that this finding underlines the importance of paying heed to questions of specificity. The trend in past research has been to speak of a global meditation response experienced to some degree by all meditators, with the same overall desirable effect. Meditation is quite likely a heterogeneous phenomenon, producing effects ranging from sleep to enlightenment, and incorporating such diverse processes as insight, desensitization, and suggestion. It is time that meditation researchers examine the question of *who* experiences *what* state and trait changes with *which* technique.

Reference Notes

1. Smith, J. C. *Problems in meditation and yoga research*. Paper presented at the International Yoga and Meditation Conference, Chicago, June 1976.
2. Otis, L. S. Meditation or simulated meditation by nonpredisposed volunteers: Some psychological changes. In E. Taub (Chair), *The psychobiology of meditation*. Symposium presented at the meeting of the American Psychological Association, Montreal, August 1973.

References

- Bergin, A. E. The evaluation of therapeutic outcomes. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis*. New York: Wiley, 1971.
- Bloomfield, H. H., Cain, M. P., & Jaffe, D. T. *TM: Discovering inner energy and overcoming stress*. New York: Delacorte Press, 1975.
- Borkovec, T. D., & Nau, S. D. Credibility of analogue therapy rationales. *Journal of Behavior Therapy and Experimental Psychiatry*, 1972, 3, 257-260.
- Cattell, R. B. *Personality and motivation structure and measurement*. New York: World Book, 1957.
- Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. *Handbook for the Sixteen Personality Factor Questionnaire*. Champaign, Ill.: Institute for Personality and Ability Testing, 1970.
- Crowne, D. P., & Marlowe, D. *The approval motive: Studies in evaluative dependence*. New York: Wiley, 1964.
- Fenz, W. D., & Epstein, S. Manifest anxiety: Unifactorial or multifactorial composition? *Perceptual and Motor Skills*, 1965, 20, 773-780.
- Fish, J. M. *Placebo therapy*. San Francisco: Jossey-Bass, 1973.
- Fitts, W. H. *Tennessee Self-Concept Scale manual*. Nashville, Tenn.: Counselor Recordings and Tests, 1965.
- Glueck, B. C., & Stroebe, C. F. Biofeedback and meditation in the treatment of psychiatric illnesses. *Comprehensive Psychiatry*, 1975, 16, 303-321.
- Kanellakos, D. P., & Ferguson, P. C. (Eds.) *The psychobiology of transcendental meditation*. Los Angeles: Maharishi International University Press, 1973.
- Maharishi Mahesh Yogi. *Transcendental meditation*. New York: Signet, 1968.
- McReynolds, W. T., Barnes, A. R., Brooks, S., & Rehagen, N. J. The role of attention-placebo influences in the efficacy of systematic desensitization. *Journal of Consulting and Clinical Psychology*, 1973, 41, 86-92.
- Orme-Johnson, D. W., Domash, L. H., & Farrow, J. T. (Eds.) *Scientific research on transcendental meditation: Collected papers*. Los Angeles: Maharishi International University Press, 1974.
- Paul, G. L. Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, 1967, 31, 109-118.

- Pratap, V. *Reactivity level in yogic and non-yogic sitting conditions* (Abstract guide of the XXth International Congress of Psychology). Tokyo: Sasaki, 1972.
- Rosen, G. M. Subjects' initial therapeutic expectancies towards systematic desensitization as a function of varied instructional sets. *Behavior Therapy*, 1975, 6, 230-237.
- Rosen, G. M. Subjects' initial therapeutic expectancies and subjects' awareness of therapeutic goals in systematic desensitization: A review. *Behavior Therapy*, 1976, 7, 14-27.
- Scheier, I. H., & Cattell, R. B. *Handbook for the Neuroticism Scale Questionnaire*. Champaign, Ill.: Institute for Personality and Ability Testing, 1961.
- Smith, J. C. Meditation as psychotherapy (Doctoral dissertation, Michigan State University, 1975).
- Dissertation Abstracts International*, 1975, 36, 3073B. (University Microfilms No. 75-27, 343). (a)
- Smith, J. C. Meditation as psychotherapy: A review of the literature. *Psychological Bulletin*, 1975, 82, 558-564. (b)
- Smith, J. C. Psychotherapeutic effects of transcendental meditation with controls for expectation of relief and daily sitting. *Journal of Consulting and Clinical Psychology*, 1976, 44, 630-637.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *Manual for the State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- White, J. *Everything you want to know about TM—Including how to do it*. New York: Pocket Books, 1976.

Received May 11, 1977 ■

Anxiety: States, Traits—Situations?

Philip C. Kendall
University of Minnesota

The present study investigated the utility of situational assessments of trait anxiety in predicting state anxiety reactions. Ninety-six male subjects pre-selected as either high or low on three measures of trait anxiety—S-R Inventory of General Trait Anxiousness (S-R GTA) Physical Danger; S-R GTA Evaluation; State-Trait Anxiety Inventory (STAI A-Trait) Anxiety Trait scale—were exposed to two experimentally induced stresses (a physical danger stress and an evaluation stress). Results indicated that the STAI A-Trait and the S-R GTA Evaluation measures correlated significantly higher with each other than either did with the S-R GTA Physical Danger measure and that both stresses produced a significant increase in state anxiety. In addition, the triple interaction of type of stress, trait level, and trait measure was, as predicted, significant. This finding indicated that high-trait-anxious subjects responded with greater state reactions when the trait measure corresponded with the type of stress. The results are discussed as support for the interaction model of anxiety and for the need to measure situational components of trait anxiety.

The state-trait model of anxiety (Spielberger, 1972) is based on the conceptual framework of transitory anxiety states (A-State) and relatively stable predispositions or traits (A-Trait). The task of state-trait researchers, according to Spielberger (1972), is to "describe and specify the characteristics of stressor stimuli that evoke differential levels of A-State in persons who differ in A-Trait" (p. 39).

The central notion of the state-trait model is that persons high in A-Trait have a greater tendency to perceive situations as dangerous or threatening than persons who are low in A-Trait, and thus they are expected to respond to threatening situations with state anxiety

elevations of greater intensity. Essentially, Spielberger views trait anxiety as the measure of anxiety proneness from which predictions of state reactions can be made.

The interaction model of anxiety (Endler, 1975) was basically derived from the rationale that had been used to develop the original S-R Inventory of Anxiousness (Endler, Hunt, & Rosenstein, 1962) and its revisions (Endler & Okada, 1975). The major point that Endler and his colleagues have made is that an appropriate assessment of anxiety must consider all sources of variability: individual differences, the responses that characterize anxiety, and the *situations* that are likely to arouse anxiety.

Recently, Endler (1975) viewed the situational component as vital for predicting state anxiety reactions. He made this position clear when he stated that

if one wants to examine the interaction of physical threat and A-Trait on state anxiety, it is necessary to assess physical danger A-Trait independent of other facets of A-Trait. (p. 161)

Within the interaction model the state-trait relationship is essentially similar to that of the state-trait model with the exception that trait

The author wishes to thank A. J. Finch, Jr., John Griffin, James F. Hooke, and Donald J. Kiesler for their helpful comments on an earlier version of this manuscript. The author also expresses thanks to M. Sue Kendall for her valuable editorial consultation.

At different times during the completion of the present paper I was affiliated with Virginia Commonwealth University and the Palo Alto Veterans Administration Hospital.

Requests for reprints should be sent to Philip C. Kendall, Department of Psychology, Elliott Hall, 75 East River Road, University of Minnesota, Minneapolis, Minnesota 55455.

anxiety is multidimensional and not unidimensional. The interaction model and the state-trait model agree that subjects high in trait anxiety will show greater state anxiety reactions under stress than will subjects low in trait anxiety, but the interaction model separates from state-trait theory in the specificity of the trait measure needed to make the differential state anxiety predictions.

The results of research using Spielberger, Gorsuch, and Lushene's (1970) State-Trait Anxiety Inventory (STAI) have indicated that individual differences in anxiety proneness (Trait Anxiety scale—A-Trait) are relatively stable and impervious to stress (Auerbach, 1973a, 1973b; Spielberger, 1972; Spielberger, Auerbach, Wadsworth, Dunn, & Taulbee, 1973), whereas the State Anxiety scale (A-State) has been found to be sensitive to various stresses (Hodges & Spielberger, 1969; Kendall, Finch, Auerbach, Hooke, & Mikulka, 1976). In addition, stressful situations of an ego-threatening nature have been found to evoke greater increases in state anxiety for high-trait-anxious than for low-trait-anxious subjects (Auerbach, 1973a; Hodges & Spielberger, 1969; O'Neil, Spielberger, & Hansen, 1969; Rappaport & Katkin, 1972) except where subjects have equal requisite skills (Kendall et al., 1976). On the other hand, in situations involving physical danger, such as threat of electric shock (Hodges & Spielberger, 1966), imminent surgery (Auerbach, Kendall, Cuttler, & Levitt, 1976; Johnson, Dabbs, & Leventhal, 1970; Spielberger et al., 1973), or films depicting physically painful accidents (Kendall et al., 1976), state anxiety reactions have been found to be unrelated to level of trait anxiety.

The S-R Inventory of General Trait Anxiousness (S-R GTA; Endler & Okada, 1975) is a self-report inventory designed to measure trait anxiety and emphasizes the importance of measuring trait anxiety in specific situations. The S-R GTA has been reported to be both a reliable measure of situational trait anxiety and one that is insensitive to momentary stress (Endler & Okada, 1975). In addition, the authors reported that the situational scales were found to be relatively independent measures.

As yet, no known study has compared the efficacy of the S-R GTA situation-specific

measures and the STAI for predicting differential state anxiety reactions. The purpose of the present study was to conduct a comparison of the models of anxiety by investigating differential state anxiety reactions in two stress situations—a physical danger stress and an evaluation stress. It was hypothesized that high-physical-danger/trait-anxious subjects (S-R GTA Physical Danger trait measure) would show greater state anxiety elevations than low-physical-danger/trait-anxious subjects for the physical danger stress, and that high-evaluation-trait (S-R GTA Evaluation trait measure) and high A-Trait (STAI) subjects would show greater state anxiety elevations than low evaluation trait and A-Trait anxious subjects for the evaluation stress. The basic hypothesis was that the trait anxiety measure corresponding to the situation would be the best predictor of the state anxiety aroused in that situation. For the S-R GTA each situational trait was expected to be predictive of that stress situation, whereas for the STAI A-Trait previous research in evaluation situations (see above), the unidimensionality of the scale (Kendall et al., 1976) and the anticipated relationship between the STAI A-Trait and S-R GTA Evaluation trait measure are indicative of accurate prediction only in evaluation stress situations. Confirmation of the hypothesis would provide support for the interaction model and would emphasize the importance of assessments of situational trait anxiety.

Method

Subjects

The subjects in the present experiment were 101 male college students selected from a pool of 173 male students (M age = 20.1) at an urban Virginia university. Only male subjects and a male experimenter participated to control for reactions characteristic of the Subject Sex \times Experimenter Sex interaction. For example, it was felt that males might protect their masculine image with a female experimenter and would subsequently report less anxiety, whereas females might seek to appear in a feminine stereotyped role with a male experimenter.

Subjects were selected based on their scores on preexperimental measures of trait anxiety. The subject selection procedure was designed to gather subject groups of high and low trait anxiety on the three preexperimental trait measures. The criterion for

Table 1
Means, Standard Deviations, and Range
of Scores of the Three Trait Measures

Trait measure	M	SD	Range
S-R GTA Physical Danger	54.09	12.19	18-75
S-R GTA Evaluation	44.34	12.32	18-73
STAI A-Trait	38.89	8.85	22-59

Note. S-R GTA = S-R Inventory of General Trait Anxiousness; STAI A-Trait = Trait Anxiety scale of the State-Trait Anxiety Inventory.

inclusion in the subject groups required scores that were in the upper 40% for the high-trait-level group and the lower 40% for the low-trait-level group. Also, to control for the other trait measures, subjects selected as high (or low) on the physical danger trait measure (S-R GTA) were not in the upper (or lower) 40% of the evaluation trait and A-Trait (STAI) scores. Likewise, subjects selected for high (or low) evaluation trait (S-R GTA) or A-Trait (STAI) groups were not in the upper (or lower) 40% of the physical danger trait scores. The means, standard deviations, and ranges of the trait scores are presented in Table 1. Three subjects were eliminated because they had previously seen the stress film, and 2 subjects were randomly eliminated to achieve an equal number of subjects in each cell. The remaining 96 subjects, 56 whites and 40 blacks, were distributed comparably across the six subject groups (one high and low group for each trait measure).

Measure of State Anxiety

State anxiety was measured using the A-State portion of the STAI. This scale consists of 20 descriptive statements that require the subjects to individually endorse on a 4-point scale (not at all, somewhat, moderately so, very much so) the degree to which each statement characterized their feelings at a particular moment in time.

Measures of Trait Anxiety

STAI A-Trait. The STAI A-Trait scale consists of 20 items that require the subjects to individually describe how they generally feel. Subjects respond to each item (e.g., "Some unimportant thought runs through my mind and bothers me") by endorsing 1, 2, 3, or 4 representing "almost never," "sometimes," "often," or "almost always."

S-R GTA. This scale consists of 15 items for each of five situations in which trait anxiety is assessed.¹ The two situations used in this study were (a) "You are in situations where you are about to or may encounter physical danger" and (b) "You are in situations where you are being evaluated by other people." The 15 items for each of these situations required responses from 1 to 5 indicating "very much" to "not at all."

Types of Stress

A physical danger and an evaluation stress were arranged to induce anxiety within an experimental laboratory setting.

Physical danger stress. This stress was a 22-minute Harvest (1970) film entitled *In the Crash*. The film presents graphic scenes of automobile crash tests at both high and low speeds. These crash tests were filmed under experimental as well as real highway circumstances. *In the Crash* has been used in previous research and has been found to be an effective stressor (Kendall et al., 1976).

Evaluation stress. The evaluation stress was a decoding task in which a word problem had to be decoded into an arithmetic solution. The problem is as follows:

$$\begin{array}{r} \text{DONALD} \\ + \text{GERALD} \\ \hline \text{ROBERT} \end{array} \quad D = 5.$$

Subjects were instructed that letters had been substituted for numbers and that the task would be to decode the letters. One part of the solution, $D = 5$, was provided. Support for the stressful nature of the task is found in previous research using similar tasks (Finch, Kendall, Montgomery, & Morris, 1975). To maximize the stress, ego-involving instructions (Spence & Spence, 1966) and a brief failure instruction (Finch et al., 1975) were included. The exact instructions and time limit are presented in the Procedure section.

Procedure

Subjects participated in groups according to one of two sequences. The sequences pertained to the order of presentation of the stress situations (physical danger then evaluation and vice versa). The two sequences were included as a counterbalanced control for order and carryover effects. The stress sequence for a given evening was decided on a restricted random basis before subjects arrived.

As subjects arrived at the laboratory room for their 7:00 p.m. appointment, they were instructed to have a seat and get comfortable. The experiment began after a 5-10 minute late grace period. Subjects were then informed by the experimenter that

I am interested in the relationship between feelings and behavior, and in order to investigate this you will be asked to fill out the "How I feel questionnaire" at certain times and later to describe some behaviors.

¹ The S-R Inventory of General Trait Anxiousness, which was published in the *Journal of Consulting and Clinical Psychology*, 1975, 43, 319-329, has only four situations. The revised version has five situations and 15 items per situation. The revised version is available from Norman S. Endler, York University, Toronto, Canada.

Following a pause, the experimenter stated that

I will be instructing you throughout your participation, so listen and you will find your job easy. If you are uncertain at any time about what you are to do, please feel free to ask.

Once everyone appeared ready, the experimenter continued by saying "What I would like you to do first is to complete the How I feel questionnaire." The STAI A-State scale was distributed with pencils, the standard instructions were read from the top of the form, and the forms were collected when all subjects had completed them.

Subjects were then instructed that they would be watching a film. They were told to "pay attention and try to get into the film, but you should not try to remember facts or details about the movie because I will not be asking questions—just watch and get the feeling." Initially, and when necessary, subjects were instructed not to talk to any of the other people during the course of the study.²

Following the film, subjects were again asked to complete the How I feel questionnaire by reporting how they had felt during the film. Subjects were next informed that the experimenter would have to rewind the film and that there were magazines available to read. The experimenter provided numerous sports and entertainment magazines (prescreened to provide an absence of stressful material, e.g., car crash pictures) and suggested that the subjects stand, stretch, relax, and look through the magazines. Subjects were reminded not to converse with each other.

When the film was rewound (approximately 10 minutes), the experimenter announced, "Now I want each of you to do something else." A pencil and a sheet of blank white paper (8½ × 11 inches) was given to each subject at his desk, and he was told to write his name across the top of the page. The experimenter then wrote the evaluation stress task on the blackboard and provided the following instructions:

This is an addition problem, only letters have been substituted for the numbers. For instance, D stands for 5. In each place where there is a D there is really a 5. Your task is to solve or decode the rest of the letters. There is a possible solution, I guarantee it. You must work on this individually, and you will be allowed a sufficient amount of time to work out the problem.

Questions were answered at this time, but no additional information was given. Finally, the experimenter provided an ego-involving comment. "This task will give me some information about your abilities that I can use to evaluate you." After 3 minutes subjects were told, "OK, everyone please stop and turn your paper over. Do not talk to your neighbors." The STAI A-State form was again distributed to the subjects while the experimenter commented:

Fill this out according to how you felt during the task. You all should have gotten the solution or at least most of it, but if you didn't, I will show you the solution at the end of tonight's projects.³

No subject solved the task in the allotted time, and all subjects should have experienced some failure and thus maximized the evaluation stress. When all STAI A-State forms were completed and collected, subjects were told that "the project is almost over, and you should sit and relax for a few minutes while I get these papers straight." After a short break subjects were asked to fill out a questionnaire not related to the present study.

Debriefing

Although the present study was rather straightforward, the debriefing explained to the subjects the actual interests of the experimenter and a brief background for the study. Subjects were also told that the experimenter had never intended to ask them to describe behavior as he had stated originally. The answer to the decoding task was presented along with a clarification of its true difficulty and an explanation that most people should not have been able to solve the problem in the allotted time. Subjects were asked not to discuss the project with other students but were told that they could talk among themselves about it. When the questions and discussion were completed, subjects were dismissed.

Only one session was held on a given evening to prevent subjects just finishing the project from passing on information to subjects just arriving. This also prevented subjects from arriving early and overhearing the debriefing session.

Results

Preexperimental Measures

Correlations of the trait measures using the original pool of 173 subjects resulted in a .38 correlation between the S-R GTA Physical Danger and Evaluation measures, a .19 correlation between the S-R GTA Physical Danger and STAI A-Trait measures, and a .52 correlation between the S-R GTA Evaluation and STAI A-Trait measures. Tests for significant differences between the correlations revealed that the S-R GTA Evaluation/STAI A-Trait correlation was significantly greater than the S-R GTA Physical Danger/STAI A-Trait

² In my previous research, induced stress tended to increase the conversion level. To prevent intersubject comparisons, subjects were instructed not to talk to each other.

³ The procedural sequence in which the evaluation stress preceded the physical danger stress was identical to that just described with the following exception: The break between stressors was covered by the experimenter stating that he had to prepare the projector for the film.

Table 2
Means and Standard Deviations of the Initial (Prestress) A-State Scores for the Subject Groups

Trait measure and level	Initial prestress A-State	
	<i>M</i>	<i>SD</i>
S-R GTA Physical Danger		
High	33.00	5.69
Low	35.62	6.48
S-R GTA Evaluation		
High	34.37	8.41
Low	30.81	7.67
STAI A-Trait		
High	37.06	7.47
Low	32.56	6.43

Note. S-R GTA = S-R Inventory of General Trait Anxiousness; STAI A-Trait = Anxiety Trait scale of the State-Trait Anxiety Inventory.

correlation, $t(167) = 4.33$, $p < .001$, and that the S-R GTA Evaluation/STAI A-Trait correlation was also significantly greater than the S-R GTA Evaluation/Physical Danger correlation, $t(167) = 1.71$, $p < .05$. These analyses indicated that the STAI A-Trait and S-R GTA Evaluation measures correlated significantly higher with each other than either did with the S-R GTA Physical Danger measure.

The means and standard deviations of the initial (prestress) A-State scores for the subject groups are presented in Table 2. An analysis of variance of the prestress scores indicated that the groups did not differ significantly, $F(5, 90) = 1.62$.

Experimental Checks

Stress. The stressful nature of the film was demonstrated by a significant t test for differences between related means of the initial and stress assessments of A-State, $t(95) = 6.08$, $p < .001$. Similarly, it was demonstrated, $t(95) = 5.50$, $p < .001$, that the evaluation task produced elevations in state anxiety. In addition, an inspection of the subjects' performance on the decoding evaluation task showed that all subjects performed at a poor level. Specifically, no one solved more than one of the letters beyond the information given.

Counterbalancing. To confirm the utility of counterbalancing the stress presentations, an analysis of variance of difference scores was conducted for subjects receiving the physical danger stress first and for those receiving it second. There was not a significant difference between the two, $F(1, 94) = 2.13$. Similarly, an analysis for the sequence of the evaluation stress revealed no significant difference, $F(1, 94) = 2.86$.

Experimental sessions. Two additional analyses were carried out to test whether there were meaningful variations in the state anxiety difference scores due to the experimental sessions. To this end, a single factor analysis of variance of A-State difference scores for the five experimental sessions was conducted. The results yielded no significant difference in physical danger or evaluation stress reactions for subjects participating in any of the sessions, $F(4, 91) = .75$, $p > .10$; $F(4, 91) = 2.45$, $p > .10$; respectively. This outcome indicated that different experimental sessions did not produce different results.

Major Hypothesis

The major hypothesis was that the change in A-State from an initial prestress period to a stress period would be greater for high-trait-level subjects than for low-trait-level subjects when the trait measure corresponded to the stressor. Thus, the A-State difference score (stress score minus initial prestress score) was the dependent measure. This hypothesis was examined via a $3 \times 2 \times 2$ analysis of variance in which type of stress was a within variable and trait level and trait measure were between variables (a three-factor mixed design with repeated measures on one factor; Winer, 1962, pp. 337-344).

The results of this analysis indicated that the main effect for trait level and the Type of Stress \times Trait Level \times Trait Measure triple interaction were significant, $F(1, 90) = 7.81$, $p < .01$, and $F(2, 90) = 9.52$, $p < .001$, respectively. The main effects of trait measure and type of stress as well as the Trait Level \times Trait Measure, Type of Stress \times Trait Level, and Type of Stress \times Trait Measure interactions were not significant. These results demonstrate that the change in state anxiety

was greater for high- than for low-trait-level subjects but was not significantly different for groups based on the trait measure or for the type of stress.

The nature of the significant triple interaction is presented in Figure 1. This illustration shows how the state anxiety difference scores varied for the trait measures, how they differed for the types of stress, and also how they varied for the high- and low-trait-level subjects. An analysis of the predicted simple effects (*t* tests) indicated that for physical danger, high-trait-level subjects were significantly greater than low-trait-level subjects on the A-State difference score under the physical danger stress, $t(30) = 3.87, p < .001$, but not significantly different under the evaluation task, $t(30) = .90, p > .10$. When high- and low-evaluation-trait subjects were compared, there was no significant difference under the physical danger stress, $t(30) = .64, p > .10$, but there was a significant difference in the evaluation task situation, $t(30) = 2.62, p < .01$, with the high-evaluation-trait subjects showing the greater A-State difference score. When high and low STAI A-Trait subjects were compared, there were no significant differences under either the physical danger stress, $t(30) = .47, p > .10$, or the evaluation stress, $t(30) = .56, p > .10$. These analyses indicated that the difference scores of the high-trait-level subjects were greater than those of the low-trait-level subjects when the situation trait measures were congruent with the stress but not when subjects were divided into high and low groups on the basis of the STAI A-Trait.

Discussion

The present study demonstrated that the STAI A-Trait and the S-R GTA Evaluation measure correlated significantly higher with each other than either did with the S-R GTA Physical Danger measure. This outcome suggests that the STAI A-Trait measure is more like the S-R GTA Evaluation measure than like the S-R GTA Physical Danger measure and appears to support previous research, which found the STAI A-Trait to be related to state anxiety in evaluation stress situations (Auerbach, 1973a; Hodges & Spiel-

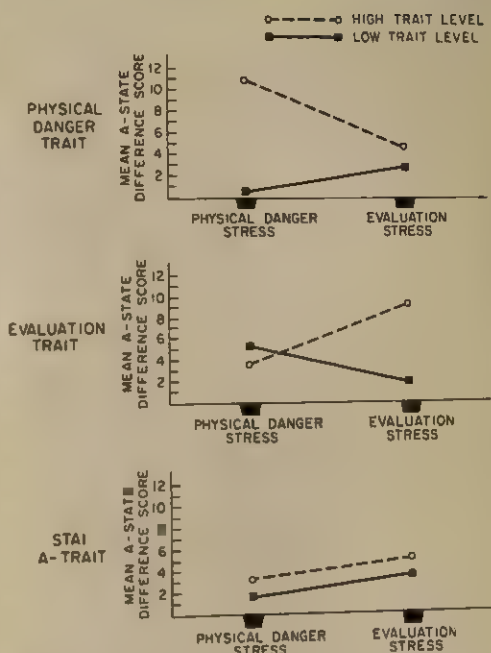


Figure 1. Mean A-State (State Anxiety scale) difference scores for the two types of stress and for subjects high and low on each of the trait measures. (A-Trait = STAI Trait Anxiety scale.)

berger, 1969). Also of interest from the correlational analyses is the relatively low relationship among the measures of trait anxiety (other than the STAI A-Trait/S-R GTA Evaluation correlation). These correlations support the notion of relatively separate anxiety traits.

The efficacy of using a situational anxiety trait measure that is congruent with the stress situation in the prediction of state anxiety reactions was demonstrated in the present study. The trait measures that were viable in the present study were two *situation-specific* measures of anxiety traits, physical danger and evaluation. When the subjects were grouped according to their physical danger trait, the high-trait-level subjects showed significant differential responsiveness to the physical danger stress but not to the evaluation stress. The opposite was true when subjects were grouped on the basis of their evaluation trait scores. On the other hand, the use of a nonsituational, unidimensional trait measure (STAI A-Trait) did not predict differential state anxiety reactions. Thus, the results of

the present study support both the need for including situations in the measurement of trait anxiety and, correspondingly, the interaction model of anxiety.

The present results also support the utility of the state-trait distinction, a distinction that has already been supported in both manipulation studies and factor-analytic work (Kendall et al., 1976; Newmark, Faschingbauer, Finch, & Kendall, 1975). However, the present investigation of "states" and "traits" suggests a clarification of their relationship: Anxiety traits are predictive of anxiety states when the trait measure is congruent with the evocative situation. It appears that there is use for the inclusion of situational trait measures.

The lack of differential state anxiety responsiveness for high and low STAI A-Trait subjects under both stresses was unexpected. Since the STAI A-Trait measure correlated more with the S-R GTA Evaluation trait measure than with the physical danger measure, since previous research has reported greater state reactions for high STAI A-Trait subjects in evaluation stresses, and since the trait measure is unidimensional, the STAI A-Trait measure was viewed as indicative of a predisposition to become anxious in evaluation situations. Based on these findings, differential state reactions for high STAI A-Trait subjects were hypothesized under the evaluation stress. However, this hypothesis was not supported in the present study.

In speculating about the reasons for the insensitivity of the STAI A-Trait measure in the evaluation stress situation, the previous findings of Kendall et al. (1976) are suggestive. In that study, in which subjects were shown to have performed equally under stress and thus were considered to have had relatively equal requisite skills, high and low A-Trait subjects reacted similarly to an examination stress. In the present experiment subjects indeed performed equally; no one decoded more than one letter beyond the given information. The present inability of the STAI A-Trait to predict differential responses in an evaluation stress could be due to the similar performance of all subjects. However, this possibility is directly contradicted by the results found when the S-R GTA Evaluation measure was used

(i.e., high- and low-evaluation-trait anxious subjects *did* respond differentially).

A more parsimonious speculation about the unexpected results using the STAI A-Trait measure concerns the measure itself. That is, whereas the STAI A-Trait measure requires subjects to indicate how they generally feel, it does not specify a situation in its assessment of the anxiety trait. The ignoring of situational specificity by the STAI A-Trait measure could account for the present findings.

More importantly, the present study provides major construct validation for the Physical Danger and Evaluation portions of the S-R GTA. Since only two situations were directly investigated, the validity of the other situation traits is yet to be examined. But both the present study and the work of Endler and Okada (1975) suggest that it would be beneficial to further examine situational measures of trait anxiety.

The outcome of the present study has implications for future research in personality in general and anxiety in particular. The generality of this finding suggests that personality trait measures should be reexamined in light of the situational dimensionality of each particular trait. This reexamination would entail investigating all the situations related to the trait, organizing these situations into *classes*, and developing an instrument for the assessment of each *situational class*. Thus, instead of continuing to assess cross-situational traits or attempting to measure situation-specific responses, researchers would potentially profit from focusing on the assessment of traits in relation to *classes* of situations.

References

- Auerbach, S. M. Effects of orienting instructions, feedback-information, and trait-anxiety level on state anxiety. *Psychological Reports*, 1973, 33, 779-786. (a)
- Auerbach, S. M. Trait-state anxiety and adjustment to surgery. *Journal of Consulting and Clinical Psychology*, 1973, 40, 264-271. (b)
- Auerbach, S. M., Kendall, P. C., Cuttler, H. F. & Levitt, N. R. Anxiety, locus of control, type of preparatory information and adjustment to dental surgery. *Journal of Consulting and Clinical Psychology*, 1976, 44, 809-818.
- Endler, N. S. A person-situation interaction model for anxiety. In C. D. Spielberger & I. G. Sarason (Eds.),

- Stress and anxiety* (Vol. 1). Washington, D.C.: Hemisphere Publications, 1975.
- Endler, N. S., Hunt, J. McV., & Rosenstein, A. J. An S-R Inventory of Anxiousness. *Psychological Monographs*, 1962, 76(17, Whole No. 536).
- Endler, N. S., & Okada, M. A multidimensional measure of trait anxiety: The S-R Inventory of General Trait Anxiousness. *Journal of Consulting and Clinical Psychology*, 1975, 43, 319-329.
- Finch, A. J., Jr., Kendall, P. C., Montgomery, L. E., & Morris, J. Effects of two types of failure on anxiety in emotionally disturbed children. *Journal of Abnormal Psychology*, 1975, 84, 583-585.
- Harvest. *In the crash*. New York: Harvest Films, 1970. (Film)
- Hodges, W. F., & Spielberger, C. D. The effects of threat of shock on heart rate for subjects who differ in manifest anxiety and fear of shock. *Psychophysiology*, 1966, 2, 287-294.
- Hodges, W. F., & Spielberger, C. D. Digit span: An indicant of trait or state anxiety? *Journal of Consulting and Clinical Psychology*, 1969, 33, 430-434.
- Johnson, J. E., Dabbs, J. M., & Leventhal, H. Psychological factors in the welfare of surgical patients. *Nursing Research*, 1970, 19, 18-29.
- Kendall, P. C., Finch, A. J., Jr., Auerbach, S. M., Hooke, J. F., & Mikulka, P. J. The State-Trait Anxiety Inventory: A systematic evaluation. *Journal of Consulting and Clinical Psychology*, 1976, 44, 406-412.
- Newmark, C. S., Faschingbauer, T. R., Finch, A. J., Jr., & Kendall, P. C. Factor analysis of the MMPI-STAI. *Journal of Clinical Psychology*, 1975, 31, 449-452.
- O'Neil, J. F., Spielberger, C. D., & Hansen, D. N. The effects of state anxiety and task difficulty on computer-assisted learning. *Journal of Educational Psychology*, 1969, 60, 343-350.
- Rappaport, H., & Katkin, E. S. Relationships among manifest anxiety, response to stress and the perception of autonomic activity. *Journal of Consulting and Clinical Psychology*, 1972, 38, 219-224.
- Spence, J. T., & Spence, K. W. The motivational components of manifest anxiety: Drive and drive stimuli. In C. D. Spielberger (Eds.), *Anxiety and behavior*. New York: Academic Press, 1966.
- Spielberger, C. D. Anxiety as an emotional state. In C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (Vol. 1). New York: Academic Press, 1972.
- Spielberger, C. D., Auerbach, S. M., Wadsworth, A. P., Dunn, T. M., & Taulbee, E. S. Emotional reactions to surgery. *Journal of Consulting and Clinical Psychology*, 1973, 40, 33-38.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. STAI manual for the State-Trait Anxiety Inventory ("Self-Evaluation Questionnaire"). Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Received May 18, 1977 ■

Comparison of Measures of Adaptive Behaviors in Preschool Children

Linda I. Garrity

Mendota Mental Health Institute and Waisman
Center on Mental Retardation and Human
Development, Madison, Wisconsin

Andria B. Servos

Department of Behavioral Disabilities,
University of Wisconsin

Six measures of adaptive behavior were compared to determine which tests and items best discriminate between behavior problem and nonproblem preschool children. Eleven nonproblem and 13 problem children, matched on age and IQ, were compared on the following tests: the Minnesota Child Development Inventory, the Classroom Adjustment Rating Scale, the Ottawa School Behavior Survey, the AML Behavior Rating Scale, the Teacher Rating Scale, and the Denver Developmental Screening Test. Problem children scored significantly lower than nonproblem children on all six measures. In addition, fine-motor items discriminated better than gross-motor, language, social, and behavioral items. The Minnesota Child Development Inventory items were the best discriminators, followed by items on the Ottawa School Behavior Survey and the Teacher Rating Scale.

The purpose of this study was to determine whether problem and nonproblem preschool children perform differently on six commonly used measures of adaptive behavior.

Although a large variety of screening instruments can be found in the literature, few studies exist that simultaneously compare a number of commonly used instruments on the same subject population. An exception is a study by Cowen, Dorr, and Orgel (1971) in which 367 5- and 6-year-olds were administered the AML Behavior Rating Scale, the Ottawa School Behavior Survey, the Teacher's Adjective Checklist, and the Teacher's Behavior Rating Scale. All four measures were significantly correlated, and the latter two discriminated between adjusted and maladjusted children.

There exists a paucity of data on preschool children despite the increasing emphasis on early screening and identification of problems. In addition, most studies comparing problem and nonproblem children do not match the

groups on both chronological and mental age. Because problem children often function at a lower level than nonproblem children of the same chronological age, the results of these studies are biased in favor of the nonproblem population, making it difficult to evaluate the instruments' effectiveness.

In the present study, the following six instruments were administered to determine which, if any, differentiate previously identified groups of problem and nonproblem children matched on both chronological and mental age.

The following are four teacher rating scales that investigate various behavioral dimensions: (a) the Teacher Rating Scale (TRS; Grossman & Levy, 1974), the AML Behavior Rating Scale (AML; Brownbridge & Van Vleet, 1969); (c) the Classroom Adjustment Rating Scale (CARS; Cowen, Note 1), which consists of the recently combined Teacher's Adjective Checklist (Cowen, Zax, Izzo, & Trost, 1966; Zax & Cowen, 1969) and the Teacher's Behavior Rating Scale (Cowen et al., 1966; Zax & Cowen, 1969); and (d) the Ottawa School Behavior Survey (SBS; Pimm & McClure, 1966).

This study is based on a master's thesis completed at the University of Wisconsin by the second author.

Requests for reprints should be sent to Linda I. Garrity, Mendota Mental Health Institute, 301 Troy Drive, Madison, Wisconsin 53704.

The Minnesota Child Development Inventory (MCDI; Ireton & Thwing, 1972) is a checklist filled out by the parent. The Denver Developmental Screening Test (DDST; Frankenburg & Dodds, 1967) is a performance test, relatively easy to administer by a trained examiner, in which the child is required to answer questions, draw pictures, and exhibit motor coordination. To date, no studies have compared problem and nonproblem children on these two tests. The present study compared the two groups by examining the entire tests as well as each individual subtest from the MCDI and the DDST.

It was hypothesized that the behavior problem children would evidence higher scores (suggesting greater maladjustment) on the above four teacher behavior rating scales, lower developmental age on the MCDI (total test and the eight subtests), and more delays (failure to perform a task passed by 90% of the children at the same chronological age) on the DDST (total test and four subtests).

Method

Subjects

Eleven nonproblem and 13 behavior problem children with no medical or physical problems participated in the study. Of the 11 nonproblem children, 7 were white and 4 were nonwhite (2 black and 2 Chicano). Seven of the nonproblem children were from middle-income families (5 white, 2 nonwhite), and 4 were from lower-income families (2 white, 2 nonwhite). Of the 13 problem children, 11 were white and 2 were nonwhite. Eleven (10 white and 1 black) of the problem children were from middle-income and 2 (1 white and 1 Chicano) were from lower-income families. The children were selected from nursery schools and day-care centers in the Madison, Wisconsin, area. Approximately 75 letters and consent forms were distributed to nursery school and day-care directors. It is not known how many of these were directly given to parents. Thirty parents volunteered participation and signed consent forms. Of these, 27 completed the study. Of the 27 original participants, 3 children were dropped from the study. The remaining 24 children were similar in both age and Peabody Picture Vocabulary Test IQ scores. The problem children had a mean age of 4.08 years ($SD = .60$) and a mean IQ of 110.53 ($SD = 11.30$). The nonproblem children had a mean age of 4.13 ($SD = .57$) and a mean IQ of 115.27 ($SD = 14.31$). The t for age = .21; the t for IQ = .85.

Children were assigned to problem and nonproblem groups on the basis of their performance in three structured situations: (a) while being administered the Peabody Picture Vocabulary Test in school by the ex-

aminer; (b) while being administered Borke's (1975) adaptation of Piaget and Inhelder's (1956) three mountains task, Flavell, Botkin, Fry, Wright, and Jarvis' (1968) block test, and the Coloured Progressive Matrices (Raven, 1956) in the clinic by the second author;¹ and (c) while participating in a mother-child interaction study (completing puzzles) in the clinic for another examiner. Specific precaution was taken to insure that the test items differentiated the children's cognitive capacity rather than attention span. The experimenter made certain through repeated demonstrations that the children understood each task before proceeding with the next. In each situation, the children were evaluated on a 6-point scale on each of two dimensions—degree of compliance and degree of on-task behavior. Children's actual scores range from 6 to 33 ($M = 19.5$, $SD = 9.6$), out of a possible range of 6 to 36. Three children whose scores were within $\frac{1}{2}$ standard deviation from the mean were excluded from the study; the sample thus consisted of 24 children.

The problem children represented both ends of the clinical spectrum. Based on the experimenter's observations, 5 of the 13 problem children showed evidence of withdrawn behavior and 4 exhibited negativistic and acting out behavior. The problems of the remaining 4 children were less clearly defined, but they could be labeled *anxious* and *immature* as evidenced by whining, crying, and so on.

Procedure

After receiving parental permission slips, the experimenter went to the child's school and administered the Peabody Picture Vocabulary Test. At that time the teacher was asked to fill out the four behavior rating scales: the CARS, the AML, the TRS, and the SBS, which required about a total of $\frac{1}{2}$ hour per child.

Mothers were invited to our center, whereupon each was asked to fill out the MCDI, which required about 20 minutes. Both mother and child were then taken to a testing room where the DDST, the Coloured Progressive Matrices, the three mountains task, and the blocks test were administered to the child.

Results

t tests were used to compare the scores of the problem and nonproblem children on the 18 measures (the six entire tests, the four DDST subtests, and the eight MCDI subtests). Significant differences between problem and nonproblem children were obtained on 16 of the 18 measures (see Table 1).

Problem children scored significantly lower than their nonproblem counterparts on seven

¹ These three tests, which were administered along with the Denver Developmental Screening Test in the clinic, are part of another study.

Table 1
Comparison of Mean Test Scores

Test	Nonproblem		Problem		<i>t</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Minnesota Child Development Inventory					
Entire test	305.54	10.14	282.15	19.20	3.63**
General Development	123.45	5.09	112.23	6.72	4.46**
Personal-Social	32.64	1.37	31.15	1.61	2.34*
Gross Motor	31.27	1.13	30.00	1.62	2.17*
Fine Motor	42.18	1.70	38.00	3.19	3.93**
Expressive Language	53.54	.05	52.15	1.66	2.75*
Comprehension-Conceptual	60.36	3.44	51.76	6.83	3.78**
Situation-Comprehension	39.72	2.37	35.92	3.36	3.14**
Self-help	33.90	2.90	31.69	3.66	1.62
Denver Developmental Screening Test					
Entire test (number of failures)	.09	.30	1.58	1.16	4.12**
Gross Motor	.09	.30	.58	.66	2.24*
Fine Motor	0	0	.58	.79	2.43*
Language	0	0	.25	.45	1.83*
Personal-Social	0	0	.16	.38	1.42
Teacher Rating Scale	34.45	11.40	50.25	9.26	3.66**
Ottawa School Behavior Survey	19.72	.64	16.30	4.15	2.70**
Classroom Adjustment Rating Scale	5.18	5.22	17.76	15.50	2.56**
Behavior Rating Scale	6.63	2.61	11.92	7.47	2.23*

Note. The degrees of freedom for the Denver Developmental Screening Test are 21; for all other tests they are 22.

* $p < .05$.

** $p < .01$.

of the eight MCDI subtests. Similar results were found on the DDST. Again, the problem children had lower developmental ages, as evidenced by a greater number of delays exhibited. The problem group had a total of 15 delays as compared to the nonproblem group, which had only 1.

A discriminant function analysis was performed using the six different tests in their entirety. The formula used to obtain a discriminant score for each subject was .058 (MCDI) + .074 (CARS) - .165 (SBS) - .091 (AML) + .088 (TRS) + 1.0 (DDST) + 15.094. Thus

Table 2
Effectiveness of the Discriminant Function Analysis

Actual	Classified as			
	Problem		Nonproblem	
	<i>n</i>	<i>p</i>	<i>n</i>	<i>p</i>
Problem	9	.69	4	.31
Nonproblem	1	.09	10	.91

to classify one subject, the scores on the six tests would be computed and multiplied by the corresponding coefficient and the constant then added. To validate the above formula, an observation was successively omitted and the data for the remaining 23 subjects were used to develop the formula to classify the 24th subject (Lachenbruch & Mickey, 1968).

The Lachenbruch and Mickey procedure was used to estimate the probability of misclassification when using the discriminant function. There were 5 errors out of 24 attempted classifications: One nonproblem child was classified as a problem child, and four problem children were classified as nonproblem. The associated probabilities are presented in Table 2.

Those items in each test that best differentiated (showed the greatest mean difference between) problem and nonproblem groups are listed in Table 3.

Discussion

This study compares the performance of problem and nonproblem children on six screen-

Table 3
*Items Within Each Test with the Greatest Mean Difference Between
 Problem and Nonproblem Groups*

Test item	Subjects		
	Nonproblem	Problem	Difference
Minnesota Child Development Inventory			
	.727	.077	.650
Reads 4 or more words	.636	.000	.636
Prints 2 or more simple words from memory	.636	.000	.636
Ties shoelaces	1.000	.385	.615
Draws recognizable pictures	.909	.308	.601
Plays table games with cards	.818	.231	.587
Prints first name	.727	.154	.573
Jumps rope	1.000	.462	.538
Colors within lines in coloring book	1.000	.462	.538
Cuts with scissors, following simple outline	.909	.385	.524
Draws picture of man/woman with at least 6 parts	.909	.385	.524
Talks in past tense correctly	.818	.308	.510
Prints a few simple words from memory			
Ottawa School Behavior Survey			
	.909	.462	.448
Says "I can't do it" before trying"	1.000	.615	.385
Doesn't complete work	1.000	.615	.385
Overstays in washroom			
AML Behavior Rating Scale			
	.114	.346	-.233
Is obstinate	.068	.250	-.182
Has difficulty learning	.205	.365	-.161
Is restless			
Classroom Adjustment Rating Scale			
	.000	.288	-.288
Poor concentration, attention span	.023	.288	-.266
Difficulty following directions	.091	.308	-.217
Defiant, obstinate, stubborn			
Teacher Rating Scale			
	.136	.521	-.384
Cooperative/uncooperative with adults	.273	.604	-.331
Good/poor self-control	.273	.583	-.311
Empathic/unempathic			
Denver Developmental Screening Test			
	.000	.292	-.292
Fine motor skills	.045	.292	-.246
Gross motor skills	.000	.125	-.125
Language skills			

ing devices. The 13 problem and 11 nonproblem children in the present study were matched on both chronological and mental age, a control procedure not ordinarily used. Assignment to problem and nonproblem groups was on the basis of the children's rated performance (on compliance/noncompliance and on-task/off-task behaviors) in three structured situations. Of the 18 measures of adaptive behavior com-

pared in the study, 16 resulted in significantly poorer performance by the problem children.

The MCDI, requiring only about 20 minutes to fill out, could easily be given to a group of parents whose children are entering nursery school. In fact, all but one of the individual MCDI subtests could be used alone as an even shorter mass screening device to identify problem children. The DDST also requires about

20 minutes, but it must be administered individually by a trained examiner. The four teacher rating scales, which also differentiated problem and nonproblem children, require no more than 5 minutes to complete and can be done by a teacher or an aide. Of the above six tests, only the MCDI and the DDST span the 6-month to 6-year range; thus they would presumably be useful with children even younger than preschool age.

The discriminant function analysis indicated that problem children would be identified correctly 69% of the time and the nonproblem children 91% of the time. It should be pointed out that the problem children in this study had no medical, physical, or previously identified behavioral problems, and all were functioning in typical day-care and nursery centers. Furthermore, problem and nonproblem children were similar on IQ, and all were functioning within the average range of intelligence. It is therefore expected that the probability of correctly identifying the more extreme problem children would be substantially higher if a random sampling procedure were used.

With regard to individual items, a possible relationship between early motor development and later behavioral and emotional problems is suggested. Of the 12 MCDI items that showed the greatest mean difference between problem and nonproblem children, 10 involved motor skills. Results on the DDST support this finding: In the nonproblem group, only one child failed or was delayed on an item that 90% of children with the same chronological age passed. In contrast, there were 15 delays (7 fine motor, 5 gross motor, 2 language, 1 personal-social) in the problem group. Thus, on both tasks nonproblem children functioned at a higher developmental level on motor skills in spite of the fact that the groups were matched on both age and IQ.

A strong relationship between motor skills and behavioral and emotional problems was also found by Rider (1973), who compared groups of "normal" and "emotionally disturbed" children 6½-12½ years old on the Purdue Perceptual-Motor Survey and the Southern California Sensory Integration Tests. The emotionally disturbed children exhibited significantly more abnormal reflex responses than the

normal children. More specifically, the emotionally disturbed children scored significantly lower on 11 of the 18 Purdue Perceptual-Motor Survey subtests and on 12 of the 16 Southern California Sensory Integration subtests. This relationship between motor development and emotional problems needs further exploration. If motor skills were found to be a good indicant of emotional problems, this would be particularly useful in screening very young children. Early identification of problem children increases the opportunity for effective intervention and treatment.

There is clearly a need for screening devices designed to detect behavioral and emotional problems in young children. Further validation of the findings of the present study is needed, using a randomly selected population as well as younger children.

Reference Note

1. Cowen, E. L. Personal communication, May 27, 1975.

References

- Borke, H. Piaget's mountains revisited: Changes in the egocentric landscape. *Developmental Psychology*, 1975, 11, 240-243.
- Brownbridge, R., & Van Vleet, P. (Eds.) *Investments in prevention: The prevention of learning and behavior problems in young children*. San Francisco: Pace ID Center, 1969.
- Cowen, E. L., Dorr, D. A., & Orgel, A. R. Interrelations among screening measures for early detection of school dysfunction. *Psychology in the Schools*, 1971, 8, 135-139.
- Cowen, E. L., Zax, M., Izzo, L. D., & Trost, M. A. Prevention of emotional disorders in the school setting: A further investigation. *Journal of Consulting Psychology*, 1966, 30, 381-387.
- Flavell, J. H., Botkin, P. T., Fry, C. L., Wright, J. W., & Jarvis, P. E. *The development of role-taking and communication skills in children*. New York: Wiley, 1968.
- Frankenburg, W. K., & Dodds, J. B. The Denver Developmental Screening Test. *Journal of Pediatrics*, 1967, 71, 181-191.
- Grossman, B. D., & Levy, P. S. A factor analytic study of coping behavior in preschool children. *Journal of Genetic Psychology*, 1974, 124, 287-294.
- Ireton, H. R., & Thwing, E. J. *Minnesota Child Development Inventory*. Minneapolis, Minn.: Behavior Science Systems, 1972.

- Lachenbruch, P. A., & Mickey, M. R. Estimation of error rates in discriminant analyses. *Technometrics*, 1968, 10, 1-11.
- Piaget, J., & Inhelder, B. *The child's conception of space*. London: Routledge & Kegan Paul, 1956.
- Pimm, J., & McClure, G. *Behavior observations of grade one pupils*. Ottawa, Canada: Ottawa City School District, 1966.
- Raven, J. C. *Coloured Progressive Matrices*. London: Lewis, 1956.
- Rider, B. A. Perceptual-motor dysfunction in emotionally disturbed children. *American Journal of Occupational Therapy*, 1973, 26, 316-320.
- Zax, M., & Cowen, E. L. Research on early detection and prevention of emotional dysfunction in young school children. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 1). New York: Academic Press, 1969.

Received May 24, 1977 ■

Effects of Controlled Background Interference on Test Performance by Right and Left Hemiplegics

Richard E. Nemec
Rehabilitation Institute of Chicago

The general and lateralized effects of background interference on verbal and perceptual-motor functioning were studied as a function of presence and lateralization of brain damage. Thirty non-brain-damaged controls and 30 right and 30 left hemiplegics were given a word-naming task and the Bender-Gestalt Test under noninterference and background interference conditions. As hypothesized, (a) brain-damaged patients had significantly greater overall interference effects than controls; and (b) laterality effects were significant, that is, verbal interference was greatest in the left-hemisphere-damaged group; perceptual interference was greatest in the right-hemisphere-damaged group. Implications for treatment programs with such patients are discussed.

The laterality hypothesis states that the left hemisphere of the brain tends to be dominant in terms of verbal functioning, whereas the right hemisphere tends to be dominant for visual, spatial, or "nonverbal functioning." This hypothesis has been the subject of a variety of studies. Many of them have upheld the hypothesis, particularly in regard to verbal functioning (Hecaen, 1962; Milner, 1967; Penfield & Roberts, 1959; Reitan, 1955; Sperry, 1965). The hypothesized function of the right hemisphere with regard to perceptual, visual, or spatial functions has not been as clearly supported and has been subject to numerous but often contradictory studies (Hanvik & Anderson, 1950; Hirshenfang, 1960; Milner, 1967; Warrington, James, & Kinsbourne, 1966).

This study was designed as a further test of the laterality hypothesis. Verbal and perceptual interference were the independent variables used to assess the effect on verbal and perceptual performance in right-hemisphere (left hemiplegics) and left-hemisphere (right hemiplegics) brain-damaged patients and non-brain-damaged controls.

The symptom of increased distractibility as a result of brain damage, resulting in the need

for limited, structured, and repetitive stimuli is commonly reported (Strauss & Lehtinen, 1947). Cantor (1966) used this symptom of distractibility to modify the Bender-Gestalt, with the dependent variable being decrement with perceptual interference. The decrement in performance was found to be highly sensitive to organic brain damage, though the variable of laterality of lesion was not controlled. Decrement as a result of interference within the verbal area has been studied (Smith, 1966) but not applied to organic-nonorganic or laterality of lesion variables per se. Thus generalized as well as lateralized effects of distractibility on performance by brain-damaged subjects would be expected.

Specifically, this study investigates the relationship of distractibility in the verbal and perceptual areas to performance on verbal and perceptual-motor tasks as a function of brain damage and laterality of lesion. Using the findings of Cantor (1967) and Strauss and Lehtinen (1947) concerning the general effects of brain damage on distractibility and the laterality literature concerning differential functioning of the left and right hemispheres, the following predictions were made:

1. Background verbal interference within the verbal mode will cause a greater decrement in the verbal performance of organics than non-

Requests for reprints should be sent to Richard E. Nemec, Rehabilitation Institute of Chicago, 345 East Superior Street, Chicago, Illinois 60611.

brain-damaged controls regardless of site of lesion.

2. Background perceptual interference in the perceptual mode will cause a greater decrement in the perceptual performance of organics than non-brain-damaged controls regardless of site of lesion.

3. Patients with lesions in the right hemisphere will show more of a decrement in the perceptual sphere with perceptual interference than patients with lesions in the left hemisphere.

4. Patients with lesions in the left hemisphere will show more of a decrement in the verbal sphere with verbal interference than patients with lesions in the right hemisphere.

5. Perceptual interference in the verbal sphere will show less of a decrement in performance than verbal interference in the verbal sphere for all groups.

6. Verbal interference in the perceptual sphere will cause less of a decrement than perceptual interference in the perceptual sphere for all groups.

Method

Subjects

The subjects in this study consisted of 90 patients involved in an active rehabilitation program in two rehabilitation hospitals. Of the 90 subjects, 60 were diagnosed by a neurologist on admission as having suffered a cardiovascular accident resulting in varying degrees of paralysis to the side of the body opposite that of the involved hemisphere. Thirty of these patients displayed primary involvement of the right hemisphere and 30 displayed primary involvement of the left hemisphere. The control group consisted of 30 patients displaying no discernable central nervous system dysfunction on neurological and psychological evaluation. The non-brain-damaged controls were all involved in an inpatient rehabilitation program, had suffered some degree of physical impairment, and were selected from an age group similar to the cardiovascular accident patients. None of the subjects were diagnosed as psychotic or severely aphasic, and those with severe perceptual problems (unable to produce a scorable Bender) were excluded from the study. Fifty-one males and 39 females participated. Mean age was 63.8 years (range = 53-72). Mean educational level was 11.2 years (range = 6-13). Age, educational level, and sexual composition were nonsignificantly different across the three groups.

Procedure

Each patient was administered the Bender-Gestalt, first using plain paper, then using Cantor's perceptual

background interference procedure. The patients were asked to copy the designs on special paper, according to Cantor's method, and the performance was transferred, via carbon paper, to a blank sheet to aid in scoring. During the verbal portion of the task, the subjects were asked to name all the animals they could within 30 sec and then repeat the task while a tape, consisting of the examiner naming tools, was presented at a constant volume (approximately 65 dB[A]). In addition, to check the possibility of cross-modal interference (i.e., verbal interference affecting perceptual performance), each patient was readministered the Bender-Gestalt under conditions of verbal background interference (tape consisting of the names of tools). To investigate the possibility of perceptual interference affecting verbal performance, each patient was asked to repeat the word-naming task while fixating a series of flashing lights. The presentations were randomized to control for any practice effect. In accordance with Cantor's technique, non-related tasks such as sections of the Wechsler Memory Scale were presented between the experimental tasks to break the patient's "set." The Bender-Gestalt tasks were coded by a colleague, so that the experimenter knew nothing of the presence or locus of brain damage. The records were scored by the examiner using Cantor's modification of the Pascal-Suttell method (scoring reliability $r = .82$). The measure of change used was Cantor's: a difference score that results from the subtraction of the standard score from the interference score, with a positive score reflecting decrement in performance.

For the verbal task, a difference score reflecting the difference between the initial verbalizations and those occurring with interference was recorded. A positive difference score reflects a decrement, whereas a negative difference score reflects an increment.

Results

A two-way analysis of variance of the effects of verbal and perceptual interference across the three groups on the verbal (word-naming) task was performed. The significant main effect among groups indicated that the neurological groups differed significantly on decrement in verbal performance with background interference, $F(2, 87) = 13.66, p < .01$. There was a significant main effect for type of interference, $F(2, 87) = 28.91, p < .01$. Verbal interference had a greater effect than perceptual interference, as expected.

Table 1 presents the means for the difference scores of the three groups under the two conditions of interference, and Table 2 presents the results of the t tests between pairs of means. Table 2 indicates that neither perceptual nor verbal interference caused significant differential decrement in the verbal performance of the non-brain-damaged controls. As predicted,

Table 1
Verbal Difference Score Means Between Groups and Interference Mode

Neurological group	Interference mode	
	Verbal	Perceptual
Non-brain-damaged controls	.53	.63
Right	1.80	.77
Left	3.97	.80

however, verbal interference within the verbal task significantly impaired the performance of both brain-damaged groups, with the left-sided group being significantly more impaired in their performance than the right-sided group. The presentation of perceptual interference during the verbal task did not produce differences in performance among the three groups.

The results of a two-way analysis of variance on the effects of verbal and perceptual interference across the same three groups on the perceptual (Bender-Gestalt) task indicated that the neurological groups differed significantly on the variable of perceptual decrement with background interference, $F(2, 87) = 26.56$, $p < .01$. There was a significant main effect for type of interference, $F(2, 87) = 22.74$, $p < .01$. Perceptual interference had a greater effect than verbal interference. A significant interaction effect indicated that the magnitude and direction of the effects of background interference differed for different neurological groups, $F(2, 87) = 22.74$, $p < .01$.

Table 2
Summary Table of t Tests for Differences Between Means on Verbal Task

Groups compared	t
NBD (verbal) vs. NBD (perceptual)	1.62
Right (verbal) vs. right (perceptual)	3.41*
Left (verbal) vs. left (perceptual)	4.30*
NBD (verbal) vs. right (verbal)	2.74*
NBD (verbal) vs. left (verbal)	3.97*
Right (verbal) vs. left (verbal)	2.82*
NBD (perceptual) vs. right (perceptual)	.67
NBD (perceptual) vs. left (perceptual)	1.21
Right (perceptual) vs. left (perceptual)	.85

Note. NBD = non-brain-damaged controls; parentheses indicate the type of interference. $df = 58$.

* $p < .01$.

Table 3
Perceptual Difference Scores Between Groups and Interference Mode

Neurological group	Interference mode	
	Verbal	Perceptual
Non-brain-damaged controls	.57	.53
Right	2.07	20.63
Left	1.50	8.30

Table 3 presents the means of the difference scores for the three groups under the two conditions of interference. Table 4 presents the results of the t tests between pairs of means. Table 4 indicates that neither perceptual nor verbal interference caused a significant differential decrement in the perceptual performance of the non-brain-damaged controls. As predicted, perceptual interference during the perceptual performance had a significant negative effect on both brain-damaged groups, with the right-sided group doing significantly poorer than the left-sided group. The presentation of verbal interference during the perceptual task did not significantly differentiate among the performance for the three groups.

Discussion

All predictions were upheld. It has been shown that the decrement in performance in the perceptual area with perceptual interference was significantly related to brain damage

Table 4
Summary Table of t Tests for Differences Between Means on Perceptual Task

Groups compared	t
NBD (verbal) vs. NBD (perceptual)	.47
Right (verbal) vs. right (perceptual)	5.28*
Left (verbal) vs. left (perceptual)	4.39*
NBD (verbal) vs. right (verbal)	1.76
NBD (verbal) vs. left (verbal)	1.43
Right (verbal) vs. left (verbal)	1.01
NBD (perceptual) vs. right (perceptual)	6.72*
NBD (perceptual) vs. left (perceptual)	4.13*
Right (perceptual) vs. left (perceptual)	3.83*

Note. NBD = non-brain-damaged controls; parentheses indicate the type of interference. $df = 58$.

* $p < .01$.

within either hemisphere, as was verbal decrement with verbal interference. Hence, interference within either mode significantly reduced the performance of brain-damaged as compared with non-brain-damaged subjects regardless of the site of the lesion.

Interhemispheric differences within the verbal mode were in the predicted direction, with the decrement for the left-hemisphere group being significantly greater than for the right-hemisphere group. Similarly, in the perceptual mode the decrement for the right-hemisphere group was significantly greater than for the left-hemisphere group. No significant cross-modal effects were found, indicating that perceptual interference had little effect on verbal performance, nor did verbal interference significantly affect perceptual performance.

Relating the results specifically to the laterality hypothesis and the nonlocalization theory, the results clearly support both views. Strauss and Lehtinen (1948) hypothesized that the effects of brain damage tend to be general and diffuse and tend to be related more to the dis-inhibiting effect of brain damage on the lower centers of the brain, which are related specifically to distractibility. This view was given support in both the verbal and perceptual areas tested. It was shown that for the sample tested, brain damage on either side caused a greater decrement in verbal functioning with verbal interference and in perceptual functioning with perceptual interference than for the non-brain-damaged controls. There were, however, findings that did argue for the laterality hypothesis; that is, the left-hemisphere-damaged group was significantly more distractible than the right-hemisphere-damaged group on a verbal task with verbal interference and vice versa for the effects of perceptual interference on a perceptual task.

In summary, the findings generally support the laterality hypothesis, that is, left hemisphere dominant for verbal functioning and right hemisphere dominant for visual-spatial functioning. The results further indicate that brain damage has both a general and a specific effect. Distractibility increases with damage to the brain, but verbal and perceptual dis-

tractibility have a differentially greater effect on the damaged left and right hemispheres, respectively. In terms of treatment and rehabilitation, this study supports the rationale for a structured, controlled, and limited stimulation program for any patient having suffered a cardiovascular accident. An additional implication is that in setting up an ideal treatment environment, one must consider not only the level of potentially distracting environmental stimulation but also the type of stimulation with reference to the site of the lesion.

References

- Cantor, A. A background interference procedure to increase sensitivity of the Bender-Gestalt test to organic brain disorder. *Journal of Consulting Psychology*, 1966, 30, 91-97.
- Cantor, A. *A scoring manual for the background interference procedure*. Iowa City: University of Iowa Press, 1967.
- Hanvik, M. L., & Anderson, A. The effect of focal brain lesions on the productions of rotations on the Bender-Gestalt test. *Journal of Consulting Psychology*, 1950, 14, 197-198.
- Hecaen, H. *Clinical symptomatology in right and left hemispheric lesions*. Baltimore, Md.: Johns Hopkins University Press, 1962.
- Hirshenfang, S. A comparison of Bender Gestalt reproductions of right and left hemiplegic patients. *Journal of Clinical Psychology*, 1960, 16, 439-442.
- Milner, B. Brain mechanisms suggested by studies of temporal lobes. In C. F. Darley (Ed.), *Brain mechanisms underlying speech and language*. New York: Grune & Stratton, 1967.
- Penfield, W., & Roberts, L. *Speech and brain mechanisms*. Princeton, N.J.: Princeton University Press, 1959.
- Reitan, R. Certain differential effects of left and right lesions in human adults. *Journal of Comparative and Physiological Psychology*, 1955, 48, 474-477.
- Smith A. Hemispheric differences in language and visual functions. *Cortex*, 1966, 1, 109-125.
- Sperry, R. W. Observations on visual perception after disconnection of the cerebral hemispheres in man. *Brain*, 1965, 88, 221-236.
- Strauss, E., & Lehtinen, L. *Psychopathology and education of the brain-injured child*. New York: Grune & Stratton, 1947.
- Warrington, E., James, M., & Kinsbourne, M. Drawing disability in relation to laterality of cerebral lesion. *Brain*, 1966, 89, 53-62.

Received July 22, 1977 ■

The Enigma of Androgyny: Differential Implications for Males and Females?

Warren H. Jones, Mary Ellen O'C. Chernovetz, and Robert O. Hansson
University of Tulsa

A program of studies ($N = 1,404$) tested the hypothesis that psychological androgyny (i.e., a balance of masculine and feminine characteristics) permits greater behavioral flexibility and consequently leads to better adjustment. A variety of methods were used to compare androgynous with sex-typed and opposite sex-typed individuals along several attitudinal, personality, and behavioral dimensions. Contrary to expectation a pattern of findings replicated across measures of attitudes toward women's issues, gender identification, neurosis, introversion-extraversion, locus of control, self-esteem, problems with alcohol, creativity, political awareness, confidence in one's own ability, helplessness, and sexual maturity indicated that flexibility and adjustment were generally associated with masculinity rather than androgyny for both males and females. A subsequent experiment further revealed that feminine subjects, independent of gender, would prefer to become more masculine were that possible. These results are interpreted as suggesting an alternative to Bem's theory of androgyny. Additional analyses indicated few differences between the additive and the original definitions of androgyny.

A common assertion in the feminist movement and in recent psychological theory is that traditional sex roles are confining and that new roles for women will result in more rewarding options for men as well. In this context, the present research tested a wide variety of adaptability-adjustment hypotheses, derived from androgyny theory (Bem, 1974).

In an important series of articles, Sandra Bem (1974, 1975) challenged the assumption frequently found in the literature that persons who adopt a conventional masculine or feminine role are somehow "healthier." Bem argued that internalizing a culturally imposed, "appropriate" sex role may inhibit the development of a full and satisfying behavioral repertoire. By contrast, the androgynous individual

who identifies with both desirable masculine and desirable feminine characteristics is freed from such stereotypic sex role limitations and is able to more comfortably and effectively engage in both "masculine" and "feminine" behaviors across a variety of social situations. Thus, the concept of androgyny denotes a person who is flexible, socially competent, able to respond to shifting situational demands, and more complete and actualized in the sense of developing and maximizing personal potential.

Initial validation studies using the Bem Sex-Role Inventory (BSRI; Bem, 1974) generally revealed the behavior of androgynous persons to be less stereotyped or constrained by conventional sex role standards. Given a choice of activities in which they believed they would be photographed, such persons were less likely than sex-typed individuals to prefer gender-appropriate activities, and if required to publicly perform a gender-inappropriate behavior, they experienced less discomfort (Bem & Lenney, 1976). Androgynous women were less dependent and conforming in a standard conformity experiment, and androgynous men were more likely than their sex-typed counter-

Preliminary reports of this article were presented at the meeting of the Southwestern Psychological Association, Albuquerque, New Mexico, April-May, 1976.

The authors gratefully acknowledge the assistance of Ken Slade, Shasta Mead, Cathy Ingle, and Shirlynn Nichol.

Requests for reprints should be sent to Warren H. Jones, Department of Psychology, University of Tulsa, 600 South College, Tulsa, Oklahoma 74104.

parts to display "feminine playfulness" with a small kitten (Bem, 1975). Masculine men were less likely to approach and play with a human baby and showed the least sympathy for the personal problems of another person. Although feminine women showed the most sympathy, they were expressively deficient with the kitten, whereas androgynous subjects were able to perform both functions adequately (Bem, Martyna, & Watson, 1976).

Bem's theory suggested a wealth of new hypotheses concerning the greater flexibility, adaptability, social competence, and psychological health of androgynous individuals, and several pertinent empirical studies using the BSRI have now been reported. A review of these studies, however, raises a number of questions regarding (a) differential patterns of results concerning the effects of androgyny for males and females, (b) the generalizability of current demonstrations of the flexibility of androgynous individuals, and (c) the validity of the BSRI.

The first issue concerns the idea that less traditional roles will be equally rewarding for both men and women. Although a few inconsistencies have been reported, the presently available evidence suggests that androgynous females are less conventional and less constrained by sex role identification than their sex-typed counterparts. For example, androgynous females as compared to feminine females have been found to be less traditional, inhibited, or restrained regarding occupational and educational objectives, marital and childbearing preferences, sexual behavior and attitudes, willingness to discuss menstrual problems, and career versus family orientations (Allgeier, 1975; Allgeier, Note 1; Kamens & Liss-Levinson, Note 2; Brooks & Birk, Note 3; Chernovetz, Hansson, & Jones, Note 4). The data for male subjects in these studies, however, are less convincing, with most studies reporting no differences, or in some instances, results contrary to expectation.

A second major issue concerns the generalizability of currently available demonstrations of the adaptability of androgynous individuals. Bem's early validation studies were creative, but they did not adequately test a range of competencies sufficient to justify the conclusion that androgynous persons are behavior-

ally and emotionally more adaptable. Indeed, much of the presently available data do not directly address the issues of mental health and social competence. Moreover, in one study, androgynous subjects yielded not only higher self-esteem and lower general maladjustment and psychosis scores but also higher scores on a neurosis scale (Nevill, Note 5). Also, the studies of investigators other than Bem contain considerable methodological variation, thereby rendering generalization difficult. For example, some studies have homogenized the subject sample, whereas others have used all available subjects. Some have failed to include opposite sex-typed subjects or males in the design, whereas others have included them, and still others have combined males and females, or opposite sex-types individuals with conventionally sex-typed individuals.

Finally, several questions concerning the validity of the BSRI as a measure of psychological androgyny have been raised. For example, low correlations between sex of subject and endorsement of specific items of the masculinity and femininity subscales (Rothman & Bryson, Note 6), possible misuse of the *t* statistic, and a possibly overlooked significant difference concerning the desirability of masculine traits for females between male and female judges in the initial scaling sample have been reported (Strahan, 1975). Furthermore, three factor analyses of BSRI responses have been reported that show contradictory results (Wakefield, Sasek, Friedman, & Bowden, 1976; Rothman & Bryson, Note 6; Gaudreau, Note 7), and the one study that attempted to replicate Bem's original scaling procedures with a new sample generally failed to do so (Edwards & Ashworth, Note 8).

Of particular concern has been the criticism of Spence and her colleagues (Spence, Helmreich, & Stapp, 1975) regarding the inability of Bem's original subtractive scoring method to detect subjects scoring low on both masculinity and femininity. The major implication is that although such persons might be androgynous as a result of their equivalent masculinity and femininity subscores, they are probably less well adjusted due to a negative self-concept reflected in such lower scores. After an empirical examination of this criticism, Bem (1977) acknowledged that on some variables the per-

formance of undifferentiated subjects (i.e., those whose masculine and feminine scores both were below the observed median) indicated a poorer sense of self-esteem than did that of the androgynous subjects (whose masculinity and femininity scores both were above the median). Such differences, she concluded, justified restricting the designation *androgynous* to persons whose masculine and feminine scores did not differ significantly, and for whom both scores were above the observed median.

In response to these considerations, we conducted a series of investigations designed to assess the implications of psychological androgyny as measured by the BSRI for individual adaptability, adjustment, and competence and to further explore the apparently inconsistent findings concerning the implications of androgyny for males and females. These objectives were pursued by testing 16 hypotheses across five areas of psychological functioning: (a) feminist ideology and gender identification, (b) personality and adjustment, (c) intellectual competence, (d) helplessness, and (e) sexual maturity and heterosexuality. In addition, the question of undifferentiated subjects was examined by comparing, on selected variables, the male and female androgynous subjects as defined by Bem's method versus the Spence et al. method.

Feminist Ideology and Gender Identification

The first major focus concerned the relationship between sex role orientation as measured with the BSRI and identification with the feminist movement—an attempt to locate the construct of androgyny in the context of contemporary and relevant social phenomena. Contrary to expectation, BSRI scores have failed to correlate with two separate measures of attitudes toward women (Zeldow, 1976; Kamens & Liss-Levinson, Note 2). It is reasonable to assume, however, that individuals, particularly females, who conceive of themselves in a less traditional manner would be more favorably disposed toward a social movement attempting redefinition of appropriate sex role behavior. Another, and perhaps less obtrusive, index of feminist orientation could be obtained from the frequency and type of gender reference used in self-descriptions, since

two common themes of the feminist movement have been consciousness raising and challenges to manifestations of sexism in the language used to describe the sexes. Thus, it was hypothesized that subjects with less conventional sex role orientations would be (a) more sympathetic to the goals and objectives of women's liberation; (b) more inclined to refer to gender in self-descriptions, particularly females; (c) but less likely to use a diminutive or denigrating form of gender reference such as girl versus woman or boy versus man.

Personality and Adjustment

A second area of inquiry involved the relationship between androgyny and other personality constructs concerned with adaptability, coping strategies, adjustment, and self-direction. The major implication of Bem's concept of sex roles is the notion that androgynous individuals are healthier and more adaptable as a consequence of their greater behavioral flexibility. Because only limited and somewhat contradictory data have been produced to date to support this assumption, five personality-adjustment dimensions were selected to assess Bem's androgyny equals mental health hypothesis including neurosis, introversion-extraversion, locus of control, self-esteem, and problem drinking.

Neurosis is generally viewed as a dispositional index reflecting tendencies toward anxiety, avoidance, moodiness, ineffective and self-defeating coping strategies, poor social skills, and a negative self-image. Within the context of neurosis, Eysenck and Rachman (1965) defined the construct *introversion-extraversion* as a continuum ranging from vulnerability to anxiety, apathy, irritability, depression, contemplation, and aloofness to hypochondriasis, sex problems, and optimism. Locus of control refers to the expectation that one's behavior will lead to desirable goals and reinforcement. Individuals might perceive the outcome of a behavioral event as under their control (internal) or beyond their control (external). Individuals who have a generalized expectancy that is external have been found to be less successful at coping with and controlling their environment, less effective at cognitive processing and academic achievement, and less

independent and self-reliant (Phares, 1976). Self-esteem denotes an evaluation of one's own adequacy and worth. Low levels of self-esteem have been related to a variety of adjustment-related variables including anxiety, unhappiness, neuroticism, drug usage and alcoholism, lack of confidence, susceptibility to external influence, and so forth (e.g., Coppersmith, 1967; Wylie, 1965). Even though the implications of problem drinking are straightforward, it is important also to note that some theories of problem drinking and alcoholism suggest the involvement of exaggerated sex role anxiety. Thus, male problem drinkers appear to derive a sense of power and masculinity from intoxication, whereas female problem drinkers gain a greater sense of feminine adequacy (e.g., McClland, Davis, Kalin, & Wanner, 1972; Wilsnack, 1973a, 1973b).

If, as we assume, androgyny implies adaptability along a variety of dimensions and in numerous settings, then it would be expected that androgynous subjects would yield a pattern of responses on these variables reflecting a strong sense of self-worth, inner-direction, social competence, and freedom from pathology. Specifically, it was predicted that androgynous individuals would be (a) less neurotic; (b) more extraverted; (c) more internal; and (d) higher in self-esteem, (e) with fewer alcohol-related problems.

Intellectual Competence

Less conventional sex role orientations have been associated with superior intellectual functioning and competence, suggesting indirect support for Bem's adaptability hypotheses, but again these differences seem more important for females than for males (Maccoby, 1966; Maccoby & Jacklin, 1974). A more direct test of the implications of androgyny for intellectual functioning might involve relevant application of intellectual capacity, for example, creative problem solving or a sophisticated level of political awareness. Creativity is defined as the production of original but appropriate solutions to a problem and has been found to correlate with a preference for cognitive complexity, independence, intuitiveness, and self-acceptance, although typically not intelligence (Butcher, 1968). Political aware-

ness has been found to predict a wide variety of important politically related behaviors and predispositions, for example, ideology, political participation and commitment, political apathy and disenchantment, and so forth (e.g., Campbell, Converse, Miller, & Stokes, 1964; Robinson, 1967).

Recent research has suggested still another related influence on the probability of successfully accomplishing one's goals in our culture, that is, an individual's self-perception of competence in endeavors requiring assertiveness, concentration, and skill preparation. One important recent study in this area found women less willing to trust their own capacities when given an opportunity to engage in either a game of skill or a game of chance (luck) to acquire a valued prize (Deaux, White, & Faris, 1975). Although this finding seems generally attributable to the dominant sex role orientation of most males and females, passivity and lack of confidence in one's own skill might be viewed as an underlying characteristic of individuals socialized as feminine, independent of gender. An analysis of such data that included the relative influence of sex and sex-type would more completely address the question of the capacity of feminine persons of either sex to acquire sufficient aspiration, assertiveness, self-preparedness, and confidence for entry into the competitive atmosphere of employment and professional endeavor.

Specific hypotheses derived from Bem's theory were that when compared to sex-typed persons, androgynous subjects would be (a) more creative, (b) more politically aware, and (c) less likely to prefer luck over skill in an effort to acquire a goal.

Helplessness

Learned helplessness is defined as a decrement or interference in instrumental responding due to inescapable and uncontrollable aversive events and has been suggested as a model of human depression (Seligman, 1975). Since the consequent behaviors of helplessness bear resemblance to the vulnerability, conformity, lack of assertiveness, and behavioral limitations attributed to sex-typed females, sex role orientation might be expected to mediate susceptibility to the helplessness effect.

More specifically, it was hypothesized that helplessness would vary as a function of the degree of femininity of the subject, with feminine females and opposite sex-typed males showing the greatest vulnerability, whereas androgynous subjects regardless of sex and sex-typed males would not become as helpless.

Sexual Maturity and Heterosexuality

One consequence commonly attributed to unconventional sex role orientation is retarded heterosexual involvement and development. Recent studies by Allgeier (1975; Allgeier, Note 1) indicate instead that androgynous females may be in fact heterosexually precocious, whereas contradictory results were reported for androgynous males. To further clarify the relationship between sex type and behaviors related to heterosexual adjustment, subjects completed an inventory requesting retrospective data concerning their adolescent sexual and dating experiences, as well as interpersonal feelings. It was hypothesized that androgynous, as compared to sex-typed subjects, would report (a) more intimate heterosexual involvement; (b) fewer heterosexually inhibiting feelings, for example, shyness; (c) greater knowledge and awareness of sex and reproduction; and (d) fewer parental restrictions in the area of sexual behavior.

Method

Subjects

Eight separate samples (total $N = 1,404$) of general psychology college students served as subjects in the several studies to be reported, in exchange for nominal course credit.

Procedure

Students in introductory psychology classes routinely completed the BSRI. Subjects were then recruited from the resulting subject pool, and each sample was independent except as indicated below.

Bem Sex-Role Inventory

The BSRI is a self-report scale designed to measure the extent of an individual's identification with desirable masculine and feminine traits (Bem, 1974). It contains 60 personality characteristics, previously scaled as being desirable traits for males (masculine items), for females

(feminine items), or desirable for both males and females (neutral items). Subjects are asked to indicate on a 7-point scale the extent to which each characteristic is "true of them." The BSRI was originally scored by calculating the t ratio of scores in response to masculine items versus responses to feminine items. In this manner, both males and females have been classified as sex typed (feminine females and masculine males), androgynous, or opposite sex typed (masculine females and feminine males).

Bem (1974) initially identified five levels of sex typing using the BSRI: feminine ($t > 2.025$, $p < .05$), near feminine ($1 < t < 2.025$), androgynous ($-1 < t < 1$), near masculine ($-2.025 < t < -1$), and masculine ($t > -2.025$, $p < .05$). For the present series of studies, three levels of classification were used: sex typed ($t > 1.0$, in the direction of the subject's sex), androgynous ($t < 1.0$, reflecting a relatively equal endorsement of masculine and feminine characteristics), and opposite sex typed ($t > 1.0$, in the direction opposite from the subject's sex). Subjects were thus classified as masculine males (MM), androgynous males (AM), feminine males (FM), feminine females (FF), androgynous females (AF), and masculine females (MF). This system of classification reflects a decision not to discard the large number of subjects who fall in the "near" sex-typed and "near" opposite-sex-typed categories. Moreover, analyses that include the entire population distribution of sex typing and not just comparisons of the more extreme cases provide a more rigorous test of the construct and would be expected to enhance the generalizability of findings.

Using this method of classification ($1 > t > -1$), Bem reported that about 50% of her samples were conventionally sex typed, 35% were androgynous, and 15% were opposite sex typed (Bem, 1975). Similarly, for the samples reported below, 51.3% of the subjects were conventionally sex typed, 34.1% were androgynous, and 14.6% were opposite sex typed, indicating the comparability of the present subjects to Bem's samples.

As previously indicated, a controversy has arisen in the literature regarding the most appropriate procedure to be used in operationalizing the construct of androgyny. Bem's subtractive technique has been criticized for its inability to detect undifferentiated subjects, who, contrary to the original formulation, appear to be deficient in self-esteem. As an alternative, Spence et al. (1975) have suggested an additive procedure in which a fourfold classification is generated based on the observed subscore medians. Thus subjects are classified as androgynous (masculine and feminine scores above the medians); sex typed or opposite sex typed (one score above, the other below the median); and undifferentiated (both scores below the median). Bem (1977) has conceded the problem that is created by including undifferentiated subjects in the androgyny category, but she prefers the use of multiple linear regression techniques as a solution.

Much of the initial appeal of the construct of androgyny resided in the simplicity of the central theme—that a balance of desirable sex-typed characteristics would have liberating effects, reducing the confinement of convention, thereby permitting interpersonal flexibility. Since the additive model appears to add the

qualifying dimension of self-esteem to the construct of androgyny and since the use of regression techniques abandons the concept of balance altogether, the original subtractive method was retained in the present series of studies as the method of classification. However, to assess the theoretical implications of the additive model, comparisons were also made between androgynous and undifferentiated subjects for feminist ideology, adjustment and personality, and intellectual competence variables.

Feminist ideology and gender identification. To test the feminism and identity hypotheses, 155 subjects completed the Women's Liberation Ideology Scale (WLIS; Goldschmidt, Gergen, Quigley, & Gergen, 1974), a 12-item scale that measures endorsement of the "feminist" orientation on several contemporary issues, for example, abortion, equal pay for equal work, and so forth. A second sample of 163 subjects completed the "I Am" test (Kuhn & McPortland, 1954), consisting of 20 sentence completion responses to the statement *I am*, from which the frequency and type of sex referents can be determined. Inclusion of a particular reference, for example, ethnic group, is considered to indicate the salience of that reference dimension to the individual respondent (Robinson & Shaver, 1969).

Personality and adjustment. Subjects ($n = 176$) completed two widely used instruments designed to assess variables theoretically related to the adjustment and social competence implications of androgyny including neurosis, extraversion-introversion (Eysenck & Eysenck, 1963), and locus of control (Rotter, 1966). Separate samples ($n_s = 168$ and 147, respectively) completed a problem with alcohol inventory (Manson, 1965) and a measure of self-esteem (Coopersmith, 1967).

Intellectual competence. The first sample listed in the above section ($n = 176$) also was instructed to list as many current United States senators as they could within a 10-minute time limit. Questions dealing with knowledge of political figures have previously been shown to discriminate on more general dimensions of political knowledge and awareness (Robinson, 1967). A separate sample of subjects ($n = 91$) completed the Unusual Uses Test, a measure of divergent creativity (Christensen, Guilford, & Wilson, 1957). Specifically, subjects were presented with the names of six objects (e.g., an old tire, a button, etc.) and were instructed to list as many uses that the object or a part of the object might serve beyond its typical function. Responses were then scored by calculating the frequency of each suggested use for each object and by assigning the value of 1 to each totally unique response generated by each subject (i.e., not listed by any other subject) and summed across the six stimulus objects. For the luck versus skill manipulation, 191 general psychology students (92 male and 99 female) previously categorized as sex typed, androgynous, and opposite sex typed using the BSRI, were given an opportunity to volunteer to participate in one of two experiments for extra course credit. They were told that both experiments involved playing an electronic game and that although all subjects would earn minimal credit simply for participating, the total amount of credit earned would be a function of how successfully one performed the game. Success in one game, they were told, was a function of luck; the

other required skill. Subjects then chose the game they preferred to try. Their choice, to trust their skill versus luck, was the dependent variable, and the games were not actually conducted.

Learned helplessness. Subjects ($n = 156$) were tested individually, following exactly the human learned helplessness procedure of Hiroto and Seligman (1975). Masculine, androgynous, and feminine males and females were assigned randomly to three pretreatment conditions in which they were (a) given an insoluble concept learning task (helplessness group), (b) given a similar but soluble concept learning task (soluble group), or (c) were simply shown the stimulus cards without completing the task (control group). Helplessness on a subsequent problem-solving task (a series of five-letter anagrams) was then measured by latency to criterion, number correct to criterion, and trials to criterion.

Sexual maturity and heterosexuality. Subjects ($n = 136$) completed an inventory requesting retrospective self-report data concerning their adolescent sexual and dating experiences, as well as feelings and experiences that might inhibit or facilitate heterosexual activity. Item content dealt with such factors as having a steady boyfriend-girlfriend prior to age 16; kissing before age 16; sexual intercourse before age 18; learning to dance prior to age 16; popularity with the opposite sex; sensitivity to criticism; shyness; awkwardness; embarrassment; sexual curiosity; sexual worry; knowledge of sex; frequency of dating; relative importance of dating versus academic performance; parental prohibitions against dating; and parental admonitions against pregnancy.

Results

A summary of means and F tests for feminist attitudes, personality, and adjustment correlates, intellectual competence variables, and items from the sexual maturity and heterosexuality scales are presented in Table 1. Table 2 contains percentages and chi-square tests for those variables elicited in a frequency data format.

Feminist Ideology and Gender Identification

Table 1 shows that as expected, sex type was related to feminist ideology. Although there was no main effect of sex type on WLIS scores, there was a significant interaction between sex and sex type. Analysis of simple effects (t tests between sex types within sex groups) revealed differences among female subjects only, with MF indicating significantly more favorable attitudes toward women's liberation issues than either AF or FF, and a trend of AF to showing more favorable attitudes than FF ($p < .10$).

Table 1

Feminist Ideology, Personality and Adjustment, Intellectual Competence, and Sexual Maturity Variables as a Function of Sex and Sex Type

Variable	n	Group M						F ratio		
		MM	AM	FM	FF	AF	MF	Sex (A)	Type (B)	A X B
WLIS	155	59.91 _a	51.73 _a	51.44 _a	55.64 _a	59.69 _a	68.11 _b	18.97**	2.32	4.62*
Locus of control	175	7.92 _a	10.92 _b	12.12 _b	10.32 _a	11.05 _a	10.22 _a	.08	3.31*	2.98
Neurosis	176	9.80 _a	10.96 _{a,b}	14.22 _b	11.17 _a	11.00 _a	11.78 _a	.18	3.62*	1.90
Extraversion	176	13.92 _a	13.92 _a	12.11 _a	11.56 _a	12.85 _a	16.22 _b	4.13*	1.85	6.61**
Self-esteem	147	19.52 _a	19.14 _a	16.00 _b	17.32 _a	18.46 _a	19.89 _a	.20	.42	5.57**
Problem drinking	168	7.56 _a	10.52 _b	11.78 _b	7.07 _a	6.12 _a	10.14 _a	3.90*	3.89*	1.10
Political awareness	176	6.39 _a	3.41 _b	5.67 _{a,b}	2.42 _a	3.68 _b	4.00 _b	6.78**	1.22	3.18*
Creativity	91	5.28 _{a,b}	3.44 _b	7.00 _b	3.69 _a	6.14 _b	4.00 _b	.77	.71	5.79*
Sexual maturity (n = 136)										
Curiosity		3.65	4.12	4.22	3.80	3.73	3.93	.68	.90	.56
Popularity		3.00 _a	2.76 _a	2.78 _a	2.46 _a	2.93 _a	3.80 _b	1.18	2.87	5.07**
Worry		2.31	2.47	2.56	2.13	2.17	2.40	.81	.41	.04
Shyness		2.81 _a	2.59 _a	3.22 _a	3.33 _a	2.40 _b	3.13 _b	1.33	2.45	4.62*
Embarrassment		2.81 _a	3.35 _a	3.44 _a	3.69 _a	2.70 _b	2.06 _b	3.33	1.86	10.14**
Knowledge		4.15	4.35	4.11	4.00	3.87	3.93	1.83	.06	.28
Dating		2.04	2.31	2.11	1.87	2.38	2.73	.41	1.47	1.08
Grades vs. dates		3.19	2.65	2.78	3.05	3.23	3.00	.92	.37	.82
Sensitive to criticism		3.35 _a	3.59 _a	4.56 _b	4.22 _a	3.53 _b	3.67 _b	.01	2.86	7.32**
Awkwardness		2.32	2.24	2.67	2.67 _a	1.77 _a	2.53 _a	.17	3.26*	1.34

Note. Within sex groups, means with different subscripts are significantly different at the .05 level. WLIS = Women's Liberation Ideology Scale; MM = masculine males; AM = androgynous males; FM = feminine males; FF = feminine females; AF = androgynous females; and MF = masculine females.

* $p < .05$.

** $p < .01$.

Not surprisingly, all females regardless of sex type scored significantly higher than males. Thus being female, and particularly, being a less traditionally sex-typed female was related to greater endorsement of contemporary women's issues, providing construct validation for the BSRI.

Somewhat contradictory results were obtained on the I Am Test, as indicated in Table 2. As a group, females listed sex more frequently than did males, $\chi^2(1) = 4.19$, $p < .04$, suggesting greater salience of gender identification for females. Among females, however, there were no significant effects of sex type on how frequently sex was included. For males there were sex type effects, with both FM ($z = 6.01$, $p < .01$) and AM ($z = 3.44$, $p < .01$) listing gender more often than MM.

Among those subjects listing their sex, females tended to refer to themselves as a girl more frequently than males referred to them-

selves as a boy, $\chi^2(1) = 3.24$, $p < .07$. This pattern was unexpected and may suggest the internalization of at least some types of stereotypic social attitudes differentiating the sexes for females.

Despite the contradictions, these data suggest that gender is currently a more salient aspect of self-concept for females, which may reflect the recent politization of female consciousness by the feminist movement.

Personality and Adjustment

The data were analyzed in a series of 2×3 (Sex \times Sex Type) analyses of variance. Table 1 reveals significant sex differences for extraversion and problem drinking, with males scoring significantly higher on both dimensions; significant sex type differences for locus of control, neurosis, and problem drinking; and significant interactions on extraversion and self-

esteem. Analysis of simple effects indicated that with the exception of extraversion, all of the personality differences associated with sex type occurred among male subjects.

Contrary to expectation, AM showed greater externality of control, more problems with drinking, and a trend ($p < .06$) toward greater introversion than MM. Similarly, FM scored more external, more neurotic, lower in self-esteem, and had more alcohol problems when compared with MM. FM were also lower in self-esteem and more neurotic ($p < .10$) than were the AM. Contrary to prediction, these results suggest that less sex-typed males experience more rather than fewer adjustment problems. For example, the relatively dysfunctional scores for problem drinking, neurosis, and self-esteem suggest existing behavioral difficulties and a poor image of self-worth, whereas the locus of control and extraversion scores imply ongoing or future inadequacy in terms of social competence and self-direction. It should be noted that the AM scored significantly in the "less adaptable" direction on only two of these scales relative to the MM. On the other hand, in no instance did the AM score significantly "more adaptable" than the MM.

Among females, analyses of simple effects indicated that MF were more extraverted than either AF or FF. No other significant differ-

ences were observed for females, although a trend suggested that MF had more problems with alcohol than did AF ($p < .07$).

Intellectual Competence

Table 1 shows that although no main effects of sex or sex type were observed, significant interaction effects were obtained for both political awareness and creativity. AM scored lower on political awareness than did MM, and there was a similar trend ($p < .09$) for creativity. Likewise, AM scored significantly lower on creativity than the FM, whereas FM and MM did not differ in these analyses. Thus, AM clearly performed more poorly than did either sex-typed or opposite-sex-typed males on these indices of intellectual competence.

By contrast, analysis of simple effects among female subjects supported the predictions, in that both AF and MF scored higher on political awareness than did FF while not differing from one another. For creativity, the expectation was more directly confirmed, with AF yielding higher scores than FF with a similar trend ($p < .10$) in relation to MF.

Thus, the intellectual variables included in the present series of studies support Bem's theory with respect to females while directly contradicting it for males. Less traditional sex role

Table 2
Frequency Data of Gender Identification, Sexual Maturity, and Luck Versus
Skill as a Function of Sex and Sex Type

Variable	%						χ^2	
	MM	AM	FM	FF	AF	MF	Males	Females
Gender identification (155)							5.65*	1.02
Listing sex	15.2	43.5	33.3	44.4	41.2	56.2	.73	1.10
Boy-girl	.0	10.0	14.3	31.2	42.9	22.2		
Sexual maturity (136)							.95	5.80*
Boyfriend-girlfriend	33.3	41.2	22.2	30.8	43.3	66.7	1.11	6.08**
Kissing	74.1	70.6	55.6	51.3	66.7	86.7	.48	6.63**
Intercourse	44.4	47.4	33.3	23.1	36.7	60.0	.03	4.77
Dancing	59.3	58.8	55.6	73.7	93.3	86.7	3.70	2.26
Pregnancy warning	70.4	41.2	55.6	43.6	60.0	60.0	.83	4.89
Luck vs. skill (191)	26.8	22.2	15.4	51.8	35.6	20.0		

Note. Percentages for the luck versus skill data represent the proportion of each group choosing luck rather than skill. Numbers in parentheses are *ns*. MM = masculine males; AM = androgynous males; FM = feminine males; FF = feminine females; AF = androgynous females; MF = masculine females.

* $p < .05$.

** $p < .01$.

Table 3

Means by Sex and Sex Type for the Helplessness Experiment

Variable and sex type		Males			Females		
		Helpless	Soluble	Control	Helpless	Soluble	Control
<i>M</i> latency to criterion	Masculine	30.65	19.95	30.43	28.80	31.60	22.60
	Androgynous	43.88	42.86	37.73	40.88	26.08	46.81
	Feminine	42.33	27.41	27.61	33.35	26.29	35.94
Trials to criterion	Masculine	9.20	6.75	8.70	10.00	11.00	8.00
	Androgynous	12.40	10.29	12.22	12.46	8.67	14.89
	Feminine	13.33	10.00	10.14	12.23	8.54	11.18
<i>M</i> errors to criterion	Masculine	2.70	1.33	2.30	2.20	3.00	2.40
	Androgynous	6.00	3.86	4.11	5.00	1.92	5.44
	Feminine	4.67	2.00	1.71	3.23	1.54	3.27

Note. Data for mean latency to criteria are in seconds.

orientations appear to be associated with greater intellectual preparedness among females, whereas the pattern among males is more complex, with both sex-typed and opposite-sex-typed males manifesting greater intellectual competence than AM.

Analysis of the luck versus skill manipulation replicated the earlier finding of Deaux et al. (1975), with males more frequently than females selecting a game of skill rather than a game of chance to earn extra course credit in general psychology, $\chi^2(1) = 7.35$, $p < .01$. However, as indicated in Table 2, further analysis revealed no sex role interaction among males and only a trend ($p < .08$) among females. Nevertheless, the pattern of percentages was in the predicted direction for females, with slightly more than half of the FF selecting the game of chance rather than the skill task. On the other hand, only one third and one fifth of the AF and MF, respectively, chose to rely on chance rather than their own abilities. These data then support Deaux et al.'s earlier findings and suggest that they may be partially accounted for by the lack of confidence of the FF group. Furthermore, these findings are consistent with the literature suggesting timidity, lack of confidence, self-effacement, and lack of a sense of self-worth among sex-typed females.

Helplessness

The three dependent variables were analyzed with a series of $2 \times 3 \times 3$ (Sex \times Sex Type

\times Helplessness Condition) analyses of variance. Table 3 presents cell means for each of these variables. The analyses revealed significant sex type effects on mean latency of trials to criterion, $F(2, 138) = 3.35$, $p < .05$, and on mean errors to criterion, $F(2, 138) = 4.09$, $p < .02$. Also, a nearly significant trend emerged for the helplessness condition on mean errors to criterion, $F(2, 138) = 2.51$, $p < .08$. No other significant main effects or interaction terms were observed. Further analyses (i.e., planned-comparison t tests between sex types, within sex groups, and collapsing across helplessness conditions) revealed that the only significant differences that occurred were between AM and MM for both mean latency to criterion, $t(50) = 2.29$, $p < .03$, and mean errors to criterion, $t(50) = 2.25$, $p < .03$.

AM were slower than MM in terms of solving of anagrams, and they made more errors before reaching criterion. Thus there were no sex type differences in susceptibility to the helplessness manipulation. However, these data parallel that of political awareness and creativity, indicating apparent deficits for the AM in the areas of problem solving and acquired knowledge.

Sexual Maturity and Heterosexuality

The sexual maturity and heterosexuality data were analyzed with 2×3 (Sex \times Sex Type) analyses of variance and 2×3 (True-False \times Sex Type) chi-square tests depending on the nature of the data. As indicated in

Table 1, several Sex \times Sex Type interactions were observed while one main effect for sex typing emerged (for awkwardness). Generally, the results indicated support for the predictions that androgynous as compared to feminine females would report more intimate heterosexual involvement and fewer feelings of inhibition, but they lack confirmation for the expectation that androgynous females would indicate greater knowledge and awareness of sex and fewer parental restrictions. Planned-comparison analyses revealed that AF reported being less sensitive to criticism, less shy, less awkward, and less easily embarrassed than FF. Also, MF reported being more popular with the opposite sex than AF, with a similar trend toward being less easily embarrassed ($p < .07$), but MF also reported being more awkward than AF. Moreover, as indicated in Table 2, analogous patterns were observed among females in the frequency data, with more active adolescent sexual and related behavior being associated with less traditional sex role orientations. More MF than AF reported having a boyfriend before age 16 ($z = 2.30, p < .02$), kissing before age 16 ($z = 2.31, p < .02$), and sexual intercourse before age 18 ($z = 2.28, p < .02$). Similarly, more MF than FF reported having boyfriends before 16 ($z = 4.00, p < .01$), kissing before 16 ($z = 4.31, p < .01$), intercourse before 18 ($z = 4.20, p < .01$), and a trend in the direction of more frequently learning how to dance before 16 ($z = 1.70, p < .09$). AF also were more likely than FF to report dancing before 16 ($z = 3.20, p < .01$), kissing before age 16 ($z = 1.86, p < .06$), and sexual intercourse before 18 ($z = 1.77, p < .07$).

Sex type was a less powerful predictor of the sexual maturity and heterosexuality items for males, however. Beyond the tendency of FM to report greater sensitivity to criticism than either MM or AM, no significant differences were observed on any of these items except for the trend of AM to report being more easily embarrassed than MM ($p < .10$).

Taken together, the male data suggest differences only in terms of what might be called *interpersonal feelings*, with traditionally sex-typed males reporting fewer personal liabilities and limitations that might be expected to inhibit or constrain heterosexual involvement.

Table 4

Means for Ideal Sex Role Identification

Item	Sex type		
	Masculine	Androgynous	Feminine
Males			
Masculine	4.63	5.13	5.63
Feminine	4.25	4.11	3.98
Females			
Masculine	4.46	4.69	5.09
Feminine	4.30	4.40	4.47

Note. Higher means indicate greater desired change.

By contrast, less traditional females not only less frequently reported restraining personal characteristics (e.g., shyness), but they also reported greater heterosexual involvement. Thus, sex role identification appears to be a less powerful influence on the heterosexual behavior of males, as was found by Allgeier (1975). However, the present analyses also suggest that greater heterosexual involvement is probably a masculine characteristic independent of gender.

Ideal Sex Role Identification

The main purpose of the present series of studies was to examine attitudinal, personality, and adjustment implications of psychological androgyny. Analysis of data from the foregoing studies, however, failed to provide support for the hypotheses concerning androgynous males. One explanation is that Bem's theory is inaccurate, at least for males. A test of this possibility would be to measure the extent to which subjects would change their personality as reflected by the BSRI if that were possible. It might be the case that certain sex type groups are dissatisfied with their present behavior and inclinations. For example, if sex-typed subjects yielded significant desired change scores in the direction opposite to their gender, this would indicate the recognition of the limitations of conventional sex role expectations and the preference to be more androgynous, thereby indirectly supporting Bem's theory. It is not implied that subjects would conceptualize such change in the language of sex typing, but rather that feminine females might prefer to be more

Table 5
Analysis of Variance of Ideal Sex Role Identification

Source of variation	df	SS	MS	F
Between subjects	134	50.496		
Sex of subject (A)	1	.185	.185	.535
Sex type (B)	2	.429	.214	.618
A \times B	2	5.233	2.616	7.561**
Error	129	44.649	.346	
Within subjects	135	56.301		
Type of item (C)	1	21.112	21.112	89.838****
A \times C	1	2.449	2.449	10.421***
B \times C	2	.382	.191	.813
A \times B \times C	2	2.092	1.046	4.451*
Error	129	30.266	.235	
Total	269	106.797		

* $p = .0135$.

** $p = .008$.

*** $p = .0016$.

**** $p = .0001$.

assertive and masculine males might desire greater tenderness.

Thus, the heterosexuality sample ($n = 136$) also completed the BSRI for a second time with a new set of instructions. For the second administration, subjects were asked to indicate the extent to which they would prefer to have more, less, or remain the same on each of the traits and qualities of the BSRI. Subjects indicated their responses on a 7-point scale, with 1 = less, 7 = more, and 4 = no change desired.

Responses to the desired change instructions for the BSRI were summed for masculine and feminine items separately for each subject, and means (by sex type) are presented in Table 4. These scores were then analyzed in a $2 \times 3 \times 2$ (Sex \times Sex Type \times Masculine vs. Feminine Items) analysis of variance with repeated measures on one factor. As indicated in Table 5, the strongest effect occurred for the type of item, with subjects indicating greater desired increases on masculine than feminine items. The Sex \times Sex Type interaction suggested that greater change was desired by opposite-sex-typed males and by sex-typed females. The Sex \times Items interaction revealed a tendency for males to desire relatively greater increases in masculinity and relatively lesser increases in femininity than did the female subjects.

Analyses of simple effects indicated that AM desired greater increases on masculine items

than did MM, $t(40) = 3.25$, $p < .01$; a similar difference between FM and MM, $t(33) = 4.41$, $p < .01$; and that FM desired greater increases on masculine traits than did AM, $t(23) = 2.20$, $p < .05$. For feminine items there were no differences among males except for a tendency of FM to desire less increase than did the MM, $t(33) = 1.83$, $p < .10$. Among female subjects, FF desired significantly greater increases on the masculine items than did AF, $t(67) = 3.53$, $p < .01$; FF showed a similar preference for increases on the masculine traits compared to MF, $t(52) = 4.78$, $p < .01$, but the difference on masculine items between AF and MF was not significant. Also there were no significant differences or trends among females on desired changes in femininity.

Thus, two important conclusions can be drawn from the ideal sex role identification data. First, both males and females indicated a strong preference for changes in the direction of masculinity as defined by the BSRI. Again, it should be noted that such desired change need not, and probably is not, conceptualized by the subjects as a desire to become more masculine. Rather, subjects indicated a strong desire to increase their capacity to behave in an instrumental fashion, that is, more assertively, more decisively, and so forth. It is interesting that the desired change in the direction of masculinity does not occur at the expense of the items measuring femininity, which might be inter-

preted as a desire to become more androgynous. However, the failure of masculine subjects to indicate the desired increases in femininity raises doubt concerning such an interpretation. With the possible exception of FM, most subjects indicated relatively little average desired change in either direction for the feminine items, suggesting a sense of satisfaction or disinterest in the areas of nurturance, emotionality, and so forth. Second, those subjects who were least masculine on the original administration of the BSRI wanted the greatest degree of increase on the masculine items. In other words, the traits that comprise the Masculinity scale of the BSRI are very important personal qualities that are highly desirable to those subjects who indicate that they do not possess them.

Androgyny and undifferentiated subjects. To assess the effect of the Spence et al. (1975) additive model, data of subjects classified as androgynous and undifferentiated were further analyzed, as indicated in Table 6. The eight variables included were selected for further analysis because of the communality among these variables and because these are the kind of data that have been used to address the issue of undifferentiated subjects. Initially, medians (combined across male and female subjects and all samples) were established for both the masculinity ($Mdn = 4.9$) and femininity ($Mdn = 5.0$) subscores. The left-hand portion of Table 6 compares those subjects originally classified as androgynous by the subtractive method who were either above both medians (high-high) and subjects who were below both medians (low-low). As can be seen, only one comparison yielded a significant difference, with low-low androgynous males manifesting significantly greater problems with alcohol than their high-high counterparts. The right-hand portion of the table reflects means and significance tests between subjects classified as androgynous versus subjects classified as undifferentiated by the additive method. It should be noted that this method not only involves partialing out high-high versus low-low androgynous subjects (with high-low and low-high subjects being assigned to sex-typed or opposite-sex-typed groups depending on gender), but it also includes high-high and low-low subjects who were originally classified as sex typed or opposite sex typed. Again, as

Table 6
Comparison of the Subtractive and Additive Methods of Classifying Androgynous Subjects

Variable	Subtractive Method								Additive Method							
	Males				Females				Males				Females			
	df	HH	LL	t	df	HH	LL	t	df	A	U	t	df	A	U	t
WLIS	16	49.28	54.45	.99	21	57.19	62.43	1.16	35	50.21	54.61	1.18	36	57.65	61.25	1.07
Self-esteem	16	19.28	19.82	.51	20	18.60	17.86	.37	35	18.86	17.70	.76	35	19.72	16.92	1.97*
Problem drinking	19	8.00	13.18	2.09**	27	6.33	5.70	.38	35	8.00	11.72	1.83*	36	7.33	5.52	1.11
Extraversion	21	11.83	11.47	.20	27	12.78	11.27	1.04	39	12.93	11.46	1.14	39	13.31	12.23	1.16
Neurosis	21	8.00	11.65	1.49	27	9.83	12.00	1.36	39	10.07	11.69	1.04	39	10.68	11.92	1.08
Locus of control	20	8.00	11.76	1.39	28	10.28	11.92	1.03	36	8.17	11.73	2.24**	60	10.20	10.70	.50
Political awareness	22	2.17	3.35	1.02	28	3.44	4.58	.95	39	4.00	4.77	.51	61	3.64	3.67	.04
Creativity	12	2.50	3.50	.76	8	5.99	5.33	.34	24	5.14	4.00	.83	22	5.25	3.67	1.36

Note. WLIS = Women's Liberation Ideology Scale. Means are presented for HH = high-high; LL = low-low; A = androgynous; and U = undifferentiated subjects.

* $p < .10$.

** $p < .05$.

indicated, only one comparison produced significant differences (male undifferentiated subjects were more external) and two nonsignificant trends.

Discussion

The initial purpose of the present series of studies was to test hypotheses derived from Bem's theory of androgyny with respect to conventionality, adaptability, social competence, and adjustment. Only partial support was obtained, however, and in two important respects, data patterns systematically contradicted predictions from Bem's theory. First, the androgyny equals adaptability hypothesis seems not to hold for males. In most instances androgynous males scored in the less adaptive direction than masculine males, and frequently these differences were significant. In no case were androgynous males found to be significantly more adaptive, flexible, or competent than masculine males. Moreover, the failure of androgynous males to yield scores suggesting more adaptive or effective coping ability occurred across a wide variety of personality, adjustment, and intellectual variables, for example, locus of control, anagram solution, alcohol problems, creativity, political awareness, and so forth. It seems reasonable to conclude that these differences derive from basic psychological processes rather than the specific tasks and instruments selected for inclusion in these studies. Also, with only two exceptions, the dispositional tendencies of feminine males appeared to be even less adaptive. Again, as compared to masculine males, feminine male subjects were less secure and flexible in numerous areas such as self-esteem, problems with alcohol, sensitivity to criticism, neurotic conflict, locus of control, and so on. Only on the intellectual dimensions of creativity and political awareness did feminine males manifest socially desirable characteristics.

Thus, a pattern emerged in which masculine males can be described as more competent and confident on numerous dimensions, whereas less traditionally sex-typed males are generally more limited and restricted, less effective and more vulnerable to influence, less sure of themselves, and perhaps even less well adjusted. Not surprisingly then, when asked to indicate

their preference for change on BSRI items, feminine and androgynous males preferred to become more masculine, whereas masculine males indicated relatively little desire to change. It is also important to note that with the exception of intellectual functioning measures, the relationship between sex typing and adaptability was generally linear among males.

Second, although greater support for Bem's formulations was obtained with female subjects, in one important aspect, there was a contradiction here as well. In support of the theory are the findings that androgynous females were less conventional, more outgoing, politically aware, creative, heterosexually active, and less awkward, shy, sensitive to criticism, and so on, than were feminine females. However, the inclusion of opposite-sex-typed females, revealed that masculine females are even more feminist in their attitudes, more politically aware, more extraverted, more popular with the opposite sex, more heterosexually involved, and so on. Stated more succinctly, the more masculine in orientation, the more adaptive, competent, and secure the female subject was. Minor exceptions to this pattern were detected, such as more drinking problems and a greater sense of awkwardness for the masculine females. However, the weight of the evidence favors the above conclusion. Moreover, when given the opportunity to indicate desired change in relation to BSRI items, it was the feminine females who expressed the greatest desire to change in the direction of masculinity, with less change indicated by androgynous females and least by masculine females. As was the case with males, the less masculine the female, the more desirable increased masculinity became.

It should be noted that the precise manner in which the present findings fail to support Bem's theory is somewhat subtle. Bem initially suggested that greater adaptability of the androgynous person can best be detected cross-situationally. That is, although she did not claim that androgynous individuals would always be the most adaptable in a single situation (e.g., least conforming, most responsive to emotional needs, etc.), she did argue that when several situations are taken into consideration, they will yield the most adaptive average pattern of behavior. As regards the first part of this argument, the present data suggest that Bem is

primarily correct, in that only for creativity and political awareness among male subjects was an androgynous group found to be significantly less adaptable or competent than their sex-typed and opposite-sex-typed counterparts. However, the notion that androgynous subjects would yield the most desirable pattern of responses across several situations is directly contradicted by the present data, in that sex-typed males and opposite-sex-typed females, with very few exceptions, showed the most flexible and competent pattern of responses.

The most succinct description of the present findings is that the more adaptive, flexible, unconventional, and competent patterns of responding occurred among more masculine subjects, independent of their gender. It is possible, therefore, that the various consequences typically attributed to sex typing might be better conceptualized as a function of masculinity or the constellation of traits that the BSRI and other inventories define as masculinity. In other words, it appears that general adaptability varies as a direct linear function of a relative mix of traits dominated by such factors as assertiveness, decisiveness, and intellectuality, as opposed to nurturance, responsiveness, and emotionality. In one regard this is not surprising, since the items that comprise the masculinity subscale have the underlying commonality of being instrumental in nature, that is, the ability to effectively and efficiently accomplish objectives. Similarly, most of the variables examined involve stereotypically masculine endeavors. The social appropriateness of such tendencies for males has long been recognized in the literature of sex typing (e.g., Broverman, Broverman, Clarkson, Rosenkrantz, & Vogel, 1970; Maccoby & Jacklin, 1974). What was unanticipated was that females who completely violated societal sex role expectations appear to be happier, more competent, and more adaptive than either androgynous or sex-typed females.

Although beyond the scope of the present study, the reason for this effect probably derives from a contingent relationship between the manifestation of instrumental behaviors and the application of various social rewards such as acceptance, approval, esteem, deference, and the like. If this is so, it would explain the tendency of subjects to judge the cross-sex

behavior of males (i.e., feminine behavior) more harshly than the cross-sex behavior of females (i.e., masculine behavior; e.g., Feinman, 1974). It may also have implications for many of the sex stereotyping studies (e.g., Broverman et al., 1970), in that what is being devalued in society is perhaps not female gender but, rather, feminine behaviors.

Thus the important issue becomes not whether one has internalized the traits and behaviors appropriate to one's gender but the extent to which one has assimilated the tendencies most highly valued by society. Some theorists (e.g., Bakan, 1966) have proposed a continuum ranging from the agentic role (i.e., a dominance of instrumentality, rationality, strength, assertiveness, masculinity, etc.) to what is termed *communality* (i.e., femininity, nurturance, emotionality, expressiveness, etc.). In a society that prefers the former to the latter, it becomes reasonable to conclude that individuals high in agentic tendencies will not only be more successful within the context of such a society's values, but such persons will feel more confident due to a history of differential application of social rewards.

Conceptualizing sex role phenomena in this manner raises several intriguing questions concerning the role of both men and women who are feminine in orientation in a society that prefers and rewards masculinity; the appropriateness of currently developing clinical techniques that attempt the "androgynization" of both sex-typed males and females (e.g., Kaplan & Bean, 1976); and the long-term implications for a society that rewards agency, perhaps to the exclusion or detriment of communality.

Regarding the utility of defining androgyny according to the additive model, the present results are inconclusive. As Bem has noted, not all comparisons need to be significant to substantiate the problem of undifferentiated subjects. However, the weight of evidence in the present series of studies suggests that the majority (30 of 32) of subtractively determined androgyny means were not significantly changed by inclusion of the undifferentiated subjects.

Moreover, the Spence et al. (1975) method raises several conceptual issues that have not as yet been fully addressed in the literature. For example, the additive model tends to ob-

scure relative differences between masculinity and femininity that in the present studies were consistently related to important behaviors and dispositions for both males and females. Similarly, in some instances the additive method classifies subjects in a manner that appears to be inconsistent with the original concept of androgyny (e.g., subjects whose masculinity and femininity subscores are not significantly different but one is above and the other below the medians are classified as sex typed or opposite sex typed). Also, the additive method defines androgyny in such a way that it may be self-esteem and not androgyny that is being measured. Thus further research is needed to determine both the validity and effectiveness of the additive definition of androgyny. In particular, studies are needed that compare the subtractive and additive methods to variables that on a priori grounds would be expected to indicate sex typing or its absence.

Reference Notes

1. Allgeier, E. R. *Heterosexuality and sex typing*. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May 1976.
2. Kamens, L., & Liss-Levinson, N. *Effect of sex role on career development and aspirations*. Paper presented at the meeting of the American Psychological Association, Chicago, August-September 1975.
3. Brooks, L., & Birk, J. M. *Do androgynous models generale androgynous counselors?* Paper presented at the meeting of the American Psychological Association, Chicago, August-September 1975.
4. Chernovetz, M. E., Hansson, R. O., & Jones, W. H. *Sex typing and sex-related inhibition*. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May 1977.
5. Nevill, D. D. *Sex roles and personality correlates*. Paper presented at the meeting of the American Psychological Association, Chicago, May 1975.
6. Rothman, M., & Bryson, J. B. *On the construct invalidity of the Bem Sex-Role Inventory*. Paper presented at the meeting of the Western Psychological Association, Los Angeles, April 1976.
7. Gaudreau, P. *Bem Sex-Role Inventory validation study*. Paper presented at the meeting of the American Psychological Association, Chicago, August-September 1975.
8. Edwards, A. L., & Ashworth, C. D. *A cross-validation of item selection for the Bem Sex-Role Inventory*. Paper presented at the meeting of the Western Psychological Association, Los Angeles, April 1976.

References

- Allgeier, E. R. Beyond sowing and growing: The relationship of sex typing to socialization, family plans, and future orientation. *Journal of Applied Social Psychology*, 1975, 5, 217-226.
- Bakan, D. *The duality of human existence*. Chicago: Rand McNally, 1966.
- Bem, S. L. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1974, 42, 155-162.
- Bem, S. L. Sex role adaptability: One consequence of psychological androgyny. *Journal of Personality and Social Psychology*, 1975, 31, 634-643.
- Bem, S. L. On the utility of alternative procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1977, 45, 196-205.
- Bem, S. L., & Lenney, E. Sex typing and the avoidance of cross-sex behavior. *Journal of Personality and Social Psychology*, 1976, 33, 48-54.
- Bem, S. L., Martyna, W., & Watson, C. Sex typing and androgyny: Further explorations of the expressive domain. *Journal of Personality and Social Psychology*, 1976, 34, 1016-1023.
- Broverman, D. L., Broverman, D. N., Clarkson, F. E., Rosenkrantz, P., & Vogel, S. R. Sex-role stereotypes and clinical judgments of mental health. *Journal of Consulting and Clinical Psychology*, 1970, 34, 1-7.
- Butcher, H. J. *Human intelligence: Its nature and assessment*. London: Methuen, 1968.
- Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. *The American voter*. New York: Wiley, 1964.
- Christensen, P. R., Guilford, J. P., & Wilson, R. C. Relations of creative responses to working time and instructions. *Journal of Experimental Psychology*, 1957, 53, 82-88.
- Coppersmith, S. *The antecedents of self-esteem*. San Francisco: Freeman, 1967.
- Deaux, K., White, L., & Farris, E. Skill versus luck: Field and laboratory studies of male and female preferences. *Journal of Personality and Social Psychology*, 1975, 32, 629-636.
- Eysenck, H. J., & Eysenck, S. B. C. *Manual for the Eysenck Personality Inventory*. San Diego, Calif.: Educational and Industrial Testing Service, 1963.
- Eysenck, H. J., & Rachman, S. *The causes and cures of neuroses*. San Diego, Calif.: Knapp, 1965.
- Feinman, S. Approval of cross-sex role behavior. *Psychological Reports*, 1974, 35, 643-648.
- Goldschmidt, J., Gergen, M. M., Quigley, K., & Gergen, R. J. The women's liberation movement: Attitudes and action. *Journal of Personality*, 1974, 42, 606-617.
- Hiroto, D. S., & Seligman, M. E. P. Generality of learned helplessness in man. *Journal of Personality and Social Psychology*, 1975, 31, 311-317.
- Kaplan, A. G., & Bean, J. P. (Eds.) *Beyond sex-role stereotypes: Readings toward a psychology of androgyny*. Boston: Little, Brown, 1976.
- Kuhn, M., & McPortland, S. An empirical investigation of self-attitudes. *American Sociological Review*, 1954, 19, 58-76.
- Maccoby, E. E. Sex differences in intellectual functioning. In E. E. Maccoby (Ed.), *The development of sex*

- differences. Stanford, Calif.: Stanford University Press, 1966.
- Maccoby, E. E., & Jacklin, C. N. *The psychology of sex differences*. Stanford, Calif.: Stanford University Press, 1974.
- Manson, M. P. *The Alcadé Test: Manual of directions and norms*. Los Angeles: Western Psychological Services, 1965.
- McCelland, D. C., Davis, W. N., Kalin, R., & Wanner, E. *The drinking man*. New York: Free Press, 1972.
- Phares, E. J. *Locus of control in personality*. Morristown, N.J.: General Learning Press, 1976.
- Robinson, J. P. *Public information about world affairs*. Ann Arbor, Mich.: Institute for Social Research, 1967.
- Robinson, J. P., & Shaver, P. R. *Measures for social psychological attitudes*. Ann Arbor, Mich.: Institute for Social Research, 1969.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1 Whole No. 609).
- Seligman, M. E. P. *Helplessness*. San Francisco: Freeman, 1975.
- Spence, J. T., Helmreich, R., & Stapp, J. Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 1975, 32, 29-39.
- Strahan, R. F. Remarks on Bem's measurement of psychological androgyny: Alternative methods and a supplementary analysis. *Journal of Consulting and Clinical Psychology*, 1975, 43, 568-571.
- Wakefield, J. A., Sasek, J., Friedman, A. F., & Bowden, J. D. Androgyny and other measures of masculinity-femininity. *Journal of Consulting and Clinical Psychology*, 1976, 44, 766-770.
- Wilsnack, S. C. Femininity by the bottle. *Psychology Today*, 1973, 6, 39-43. (a)
- Wilsnack, S. C. Sex role identity in female alcoholism. *Journal of Abnormal Psychology*, 1973, 82, 253-261. (b)
- Wylie, R. Self-ratings, level of ideal-self ratings and defensiveness. *Psychological Reports*, 1965, 16, 135-150.
- Zeldow, P. B. Psychological androgyny and attitudes toward feminism. *Journal of Consulting and Clinical Psychology*, 1976, 44, 150.

Received August 16, 1977 ■

During Training and Posttraining Effects of Live and Taped Extended Progressive Relaxation, Self-relaxation, and Electromyogram Biofeedback

Irving Beiman, Eileen Israel, and Stephen A. Johnson
University of Georgia

This study compared live and taped progressive relaxation (LR, TR), self-relaxation (SR), and electromyogram biofeedback (BF) on measures of autonomic and somatic arousal and subjective tension. Male and female respondents ($N = 40$) to an ad for therapy were evaluated in five training sessions and a posttraining assessment of self-control. During training, LR was superior to TR on reductions in physiological arousal; SR and BF were equivalent except for the superiority of SR on reductions in autonomic arousal. After training, LR was superior to the other procedures on self-control of autonomic arousal. It was concluded that LR is the treatment of choice for a variety of clinical objectives.

Although progressive relaxation training appears to be widely used in the clinical setting, research is equivocal regarding the physiological effects of the various forms of the procedure (Mathews, 1971). Using unselected, nonvolunteer female psychology students as subjects, Paul and Trimble (1970) found abbreviated live training superior to taped training on all physiological systems measured. Russell, Sipich, and Knipe (1976) also found live training superior to taped training for undergraduate females. Considering the economy and efficiency potentially afforded by taped training, it is important to determine the generalizability of these results to the clinical setting. Thus, one purpose of the present investigation was to compare the during training effects of extended live progressive relaxation to taped relaxation in a clinical population.

This investigation was supported by a small grant awarded to the first author by the Office of General Research, University of Georgia. The portions of the during training data (first three training sessions excluding biofeedback), based on a thesis by the second author, were published earlier.

Appreciation is expressed to Ernest Fruge and Lewis Humphreys, who served as therapists and to Earl Robinson, who assisted the first author by programming the computer analyses.

Requests for reprints should be sent to Irving Beiman, Department of Psychology, University of Georgia, Athens, Georgia 30602.

Other forms of relaxation training have also been recommended to therapists, including electromyogram (EMG) biofeedback (Cox, Freundlich, & Meyer, 1975; Reinking & Kohl, 1975) and self-relaxation training (Benson, 1976; Benson, Greenwood, & Klemchuk, 1975). Reinking and Kohl found EMG biofeedback superior to taped relaxation instructions for the facial muscles on reductions in forehead muscle tension in unselected psychology students. Using selected clinical subjects complaining of tension headaches, Cox et al. (1975) found live progressive relaxation training and EMG biofeedback equally superior to a medication placebo group on posttreatment resting frontalis EMG. Beary, Benson, and Klemchuk (1974), using male and female normal subjects in a one group/own control design, found a self-relaxation procedure to be effective in reducing oxygen consumption, carbon dioxide production, and respiratory rate relative to control periods.

Since a self-relaxation procedure with a therapeutic rationale has not been compared to any other of the more commonly used relaxation procedures, a second purpose of this study was to compare such a procedure to EMG biofeedback on during training reductions in physiological arousal and subjective tension. Third, since these four relaxation procedures have not previously been compared

on posttraining control of tension level, a final purpose was to evaluate their relative effectiveness in teaching clients self-control skills for the reduction of physiological arousal and subjective tension.

Method

Subjects

Participants were 19 males and 21 females chosen from respondents to local newspaper ads that solicited tense people to participate in a psychological study to alleviate their tension. On the basis of an initial clinical interview, respondents were selected who were free from acute illness, acute situational stress that could have changed during the course of therapy, and usage of psychoactive medication; had not previously received training in any of the treatment procedures to be used; were not in therapy to deal with tension; and indicated that tension was a serious problem for them. Participants were then ranked according to age ($M = 27.1$; range = 20-54) and were randomly assigned to one of four training conditions ($n = 10/\text{group}$): live progressive relaxation (LR); taped progressive relaxation (TR); self-relaxation (SR); and EMG biofeedback (BF). Female-to-male sex ratios were 5:5 for all groups with the exception of BF, which was 6:4. Each subject was scheduled for six sessions at the same time, each 3 or 4 days apart, with treatment conditions random over time of day. As a further control for physiological variations, each female received her first session at least 1 and not more than 2 weeks after the onset of her last menstrual cycle.

Apparatus

Treatment conditions were conducted in two adjoining soundproof, airconditioned, and electrically shielded experimental chambers. All physiological recordings were made on a Grass Model 7 polygraph with five 7DAC driver amplifiers. Standard preamplifiers and a Beckman electrolyte were used except as described below. Heart rate was recorded by gold-plated electrodes attached to the left wrist and leg; respiration was recorded by a thermistor placed slightly inside the nostril permitting the most air flow; integrated muscle tension was recorded from the frontalis muscle, using standard placement (Venables & Martin, 1967, p. 285) with electrodes similar to those for the electrodermal system. A Consol BSR/GSR preamplifier providing a constant current of 16 μA was used to record skin resistance level (SRL) and skin resistance response (galvanic skin response; GSR). Beckman silver-silver chloride electrodes 2 cm^2 in diameter were attached to the volar surface of the left palm and left forearm. The electrolyte was a saline solution in Unibase, as recommended by Lykken and Venables (1971). A GSR was defined as any decrease in resistance exceeding .1% of basal skin response, as recommended by Edelberg (1967). Each subject was seated in a comfortable recliner chair.

For feedback, the output from the EMG driver amplifier (integrator time constant = .2 sec) was connected to a Narco Limit Indicator (LI 300) permitting the establishment of an adjustable threshold. When the biological signal was below threshold, the LI-300 turned on a Mallory Sonalert, which provided a 1,000-cps tone. The tone volume was adjusted at the beginning of each feedback session by a variable resistance potentiometer according to the subject's preference.

Procedure

Training. For all groups, the first five sessions involved training. Upon the subject's arrival at the first session, the polygraph operator introduced himself and the therapist (in the LR condition only) and presented a general introduction to the study. The subject's particular problem with tension was then discussed, and a therapeutic rationale for the treatment was presented. The polygraph operator then attached the transducers and administered the Anxiety Differential (Husek & Alexander, 1963). This was followed by a 10-min adaptation period, during which the subject sat quietly in a semirecumbent position. The last 3 min of this period in each session served as the pretreatment basal level for all physiological measures. Training procedures were then begun (described below) with eyes closed for all subjects. The final 3 min of training served as the posttreatment assessment for all physiological measures. The posttraining Anxiety Differential was then administered, electrodes were detached, and the subject was instructed to practice his/her relaxation skills once per day. Home practice was discussed at the beginning of the second through the fifth sessions, and solutions were offered for any difficulties encountered.

Therapists in the LR and TR conditions were two male graduate students in clinical psychology. They were trained by the first author in the administration of a shortened form of the progressive relaxation training described by Bernstein and Borkovec (1973) and were experienced with the procedure. Each therapist treated one half the participants in each of the relaxation training conditions. Separate tapes were recorded by each therapist for each of the five training sessions. Taped presentation differed from live presentation only in that progression to the next muscle group was not contingent on the client's report of complete relaxation in the current muscle group. The first two sessions involved training in 16 muscle groups with tension release, the next two sessions involved 4 muscle groups with tension release, and the fifth session involved training in relaxation by recall. Thus the treatment package in this study progressed over five sessions rather than the nine sessions, as described by Bernstein and Borkovec. Training was preceded in Sessions 1, 3, and 5 by a description of tensing instructions. The actual administration of training lasted approximately 35 minutes in the first two sessions, 15 minutes in the next two, and 5 minutes in the fifth session.

BF training involved binary auditory feedback for successively lower amounts of integrated frontalis

muscle action potential. The threshold was set at the beginning of each session so that the tone was on approximately 50% of the time. When the tone had been continuously on (EMG level below threshold) for 15 sec, the threshold was lowered approximately 2.5% (1–3 μ v). If the subject did not meet the 15-sec criterion within a 2-min period, the threshold was gradually raised until the criterion was met. Thus the procedure involved gradually shaping the desired response.

In SR training the rationale emphasized the subject's potential to develop control over her/his tension level by regular practice of the relaxation response (cf. Beary et al., 1974). The client was instructed to relax as much as possible but not to go to sleep. As a control for time from pretraining to posttraining for BF, SR training also lasted 30 minutes for each training session.

Assessment of self-control. After the standard pretraining physiological assessment in the sixth session, the subject was instructed to relax as much as possible using the relaxation skills he/she had developed in the previous five training sessions. This self-control period lasted 10 minutes, with the final 3 minutes serving as the postphysiological assessment. Another Anxiety Differential and additional questionnaires (described below) were administered, electrodes were detached, and the subject was debriefed.

Assessment of trait anxiety. To assess potential changes in trait anxiety, all subjects were administered the Trait scale of the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1968) and the Multiple Affect Adjective Checklist (MACL; Zuckerman & Lubin, 1965) in the initial screening session and at the end of the sixth session.

Results

After confirming that there were no pre-treatment differences between groups on any measure in Session 1 (all $ps > .20$), the data for the first five training sessions were analyzed apart from the sixth session. The sixth session was conceptualized as an assessment of the subjects' ability to reduce their own arousal level when asked to do so without the concurrent aid of any therapeutic procedure. The first five sessions afforded an assessment of each procedure's specific potential for accomplishing, during training, in-session reductions in physiological and subjective tension.

The results for the five training sessions were analyzed separately: LR was compared to TR; and SR was compared to BF. This was because the duration of LR and TR was reduced systematically across sessions by virtue of the reduction from 16 muscle groups in the first session to relaxation by recall in the fifth session. The duration of training for SR and BF was held constant at 30 minutes

for each of the five sessions. This difference between the two pairs of treatments was implemented to enhance the external validity of results. Progressive relaxation training in the clinical setting is reduced in duration as training progresses (Bernstein & Borkover, 1973), whereas the duration of BF training is typically the same for each training session.

Physiological variables analyzed were electrodermal response (GSR frequency/min), heart rate (beats/min), respiration rate (cycles/min), and muscle tension (mean μ v/min). The data were quantified by trained raters, with inter-scoring reliability exceeding .99. After the reduced data were keypunched, all statistical analyses were performed using SOUPAC programs on the IBM 360 computer of the University of Georgia. A significance level of .05 was adopted for all statistical tests.

Assessment of Training Effects

Live versus taped progressive relaxation training. In-session changes for the five training sessions were evaluated by three-way analyses of variance (Treatment \times Session \times Pre-Post) on each of the five dependent variables. These analyses revealed significant pre-post main effects for respiration rate, heart rate, muscle tension, and the Anxiety Differential, $F_s(1, 18) = 10.31, 4.82, 13.38$ and 44.31 , respectively. Each of these main effects indicated that there was a significant reduction from pretraining to posttraining, although this was qualified by higher order interactions for all measures with the exception of respiration rate.

There were significant Treatment \times Pre-Post interactions for the measures of autonomic arousal—frequency of electrodermal response and heart rate, $F_s(1, 18) = 4.81$ and 9.93 , respectively. Pre-post change averaged across sessions for GSR frequency/min was -1.45 for LR and $+1.94$ for TR; and for heartbeats/min, -2.54 for LR and $+.46$ for TR. Because LR led to decreases in both GSR frequency and heart rate while TR led to increases, the significant interactions for these two measures indicate that live relaxation training was superior to taped training in accomplishing in-session autonomic relaxation.

The analyses revealed a significant Treatment \times Session \times Pre-Post interaction for muscle tension, $F(4, 72) = 3.21$. The same three-way interaction approached significance for the measure of subjective tension, the Anxiety Differential, $F(4, 72) = 2.35, p < .08$. Figure 1 depicts the mean change in muscle tension from pretraining to posttraining for each of the five training sessions. Analysis of the simple main effects for each session (Kirk, 1968, p. 222) indicated in the first session that TR was superior to LR and that in the second and third sessions that the treatments were equivalent, whereas in the fourth and fifth sessions, live training led to greater reductions in muscle tension than taped training, $F_s(1, 90) = 8.07, .11, .20, 4.65$, and 4.60 , respectively. For taped training there was a clear trend toward attenuated treatment effects as training progressed, whereas reductions in muscle tension generally increased across sessions with live training. Figure 2 presents the pre-post means in each session for the Anxiety Differential. Post hoc analyses were not performed on these data, because the interaction was not significant. Descriptively, LR led to greater mean reductions in subjective tension than TR in all but the fifth session, and LR subjects consistently attained a deeper mean level of subjective relaxation than TR subjects.

To summarize, live progressive relaxation training was superior to taped training in accomplishing in-session relaxation on three of the four physiological variables measured. Live training led to significant decreases in autonomic arousal (GSR frequency and heart

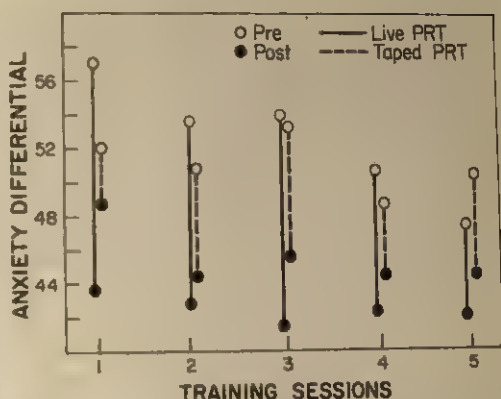


Figure 2. Pre/post Anxiety Differential means for each training session. (PRT = progressive relaxation training.)

rate) across sessions, whereas taped training resulted in increases in these two autonomic systems. Furthermore, live training was superior to taped training in reducing muscle tension, and a similar nonsignificant trend was noted for subjective tension.

EMG biofeedback versus self-relaxation training. In-session changes in the five training sessions for these two groups were evaluated by similar analyses of variance (Treatment \times Session \times Pre-Post). These analyses indicated significant pre-post reductions for heart rate, muscle tension, and the Anxiety Differential, $F_s(1, 18) = 25.26, 7.90$, and 14.03 , while the main effect for GSR approached significance, $F(1, 18) = 3.99, p < .06$.

Main effects for the two autonomic measures were qualified by significant interactions. The Treatment \times Pre-Post interaction for heart

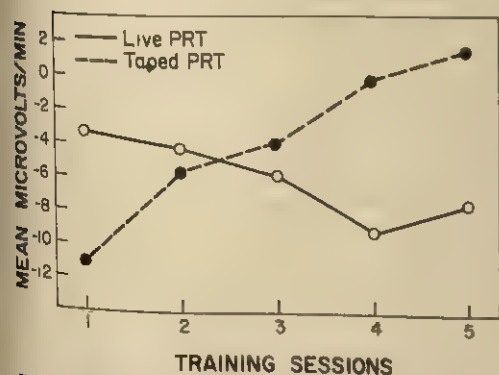


Figure 1. Mean pre to post change in muscle tension for each training session. (PRT = progressive relaxation training.)

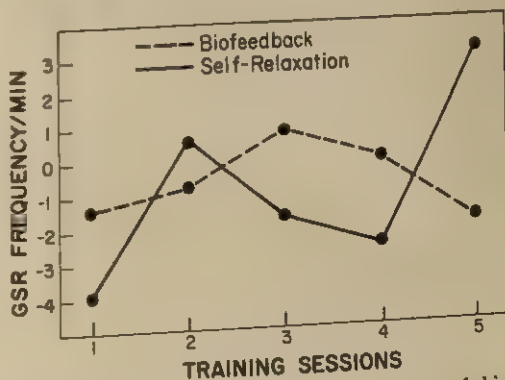


Figure 3. Mean pre to post change in frequency of skin resistance responses for each training session. (GSR = galvanic skin response.)

rate, $F(1, 18) = 4.29$, indicated that SR was superior to BF: Pre-post change averaged across training sessions in beats/min was -4.34 for SR and -1.91 for BF. The Treatment \times Sessions \times Pre-Post interaction for frequency of electrodermal response, $F(4, 72) = 4.69$, is presented in Figure 3. Analysis of simple main effects for the change scores in each training session indicated no differences ($p > .10$) between the two treatments in the first four sessions, $F_s(1, 90) = 2.35, .74, 2.58$, and 2.56 , respectively. In the fifth training session, BF was apparently superior to SR, $F(1, 90) = 8.37$. Closer examination, however, revealed that the atypically large increase from pretraining to posttraining for SR could have been a function of a low pretraining baseline (2.56 responses/min). This, in conjunction with a high pretraining baseline for BF (7.0 responses/min), could account for the significant interaction. Analysis of simple main effects confirmed, in Session 5, that BF had a higher pretraining baseline than SR, $F(1, 90) = 7.52$. Interpretation of the significant interaction is therefore seriously qualified by differential pretraining baselines in Session 5.

To summarize, both SR and BF led to significant reductions in somatic and subjective tension. Additionally, self-relaxation was differentially more effective than EMG biofeedback in reducing heart rate.

Assessment of Self-control

The sixth session data were analyzed by two-way analyses of variance (Treatment \times Pre-Post) to evaluate the participants' self-control over tonic physiological arousal and subjective tension. There were pre-post main effects for all voluntary response systems (cf. Paul, 1969): respiration rate, muscle tension, and the Anxiety Differential, $F_s(1, 36) = 19.99, 10.59$, and 17.78 , respectively. The analyses additionally indicated significant Treatment \times Pre-Post interactions for frequency of electrodermal response and respiration rate, $F_s(3, 36) = 3.26$ and 3.40 . Duncan's multiple-range test (Duncan, 1955), presented in Table 1, was applied to the mean change scores for each group to analyze the differential change from pretraining to posttraining. The ordered means for GSR frequency were LR = -2.98 ; TR = $+1.56$; BF = $+1.70$; SR = $+2.70$. The statistical comparison ($df = 36$) among all possible pairs of means indicated that live relaxation was superior to each of the other treatments, which were not different from each other. The ordered means for respiration rate were SR = -2.86 ; LR = -2.47 ; BF = $-.50$; TR = $-.36$. The multiple-range test indicated that neither self-relaxation nor live relaxation were different from each other, and both were superior to biofeedback and taped relaxation.

Table 1

Duncan's Multiple-Range Comparisons on Physiological Change Scores During Assessment of Self-control

Group	Obtained	SSR	Obtained	SSR	Obtained	SSR
	TR		BF		SR	
GSR frequency						
LR	4.54	> 4.04	4.68	> 4.25	5.68	> 4.36
TR			.14	4.64	1.14	4.25
BF					1.00	4.04
	LR		BF		TR	
Respiration rate						
SR	.79	1.84	2.36	> 1.94	2.50	> 1.90
LR			1.97	> 1.84	2.11	> 1.94
BF					.14	1.84

Note. SSR = shortest significant range; LR = live relaxation; TR = taped relaxation; BF = biofeedback; SR = self-relaxation; GSR = galvanic skin response. All comparisons in which the obtained value is greater than the SSR value are significant at the .05 level.

Assessment of Trait Anxiety

Treatment \times Pre-Post analyses of variance were performed on the ACL and Trait form of the STAI, which were administered prior to Session 1 and at the end of Session 6. There were significant pre-post main effects for both the Trait scale and the ACL, $F_s(1, 36) = 16.77$ and 19.00 . This indicated that all groups reported reductions in trait anxiety (Trait scale: pre = 49.66, post = 42.69; ACL: pre = 11.87, post = 8.58), with no differential change across groups.

Discussion

During Training Effects

The present results are consistent with the earlier results of Paul and Trimble (1970) and Russell et al. (1976): Live relaxation training was superior to non-response-contingent taped training in producing decreases in physiological arousal during training. Paul and Trimble's data were obtained from unselected nonvolunteer female psychology students, and abbreviated relaxation training was used (two sessions of training with 16 muscle groups). The present study extends those findings to older self-referred male and female clients. The fact that LR was superior to TR across all five training sessions on both measures of autonomic arousal is significant because the progression, in the present study, from 16 muscle groups to relaxation by recall is similar to actual clinical practice. The clinical context of this investigation and the emphasis on regular daily practice of relaxation skills further enhance the study's external validity. Thus the present results are readily generalizable to the clinical setting and contraindicate the use of taped training when reductions in physiological arousal are the clinical objective.

A preliminary report of the during training data involving a comparison of LR, TR, and SR in the same analysis (Israel & Beiman, 1977) yielded the conclusion that LR was superior to TR and SR on reductions in subjective tension, with no differential training effects for reductions in physiological arousal. These analyses involved only three training sessions with fewer subjects than the present analyses. Including the SR group in the

analyses apparently introduced a time confound because of the differential durations of the progressive and self-relaxation procedures. These factors, combined with the reduced power of the earlier analyses (Israel & Beiman, 1977), account for the differential results of the two studies. The present, more precise analyses are more appropriate for the evaluation of differential effects of live and taped modes of extended training.

Paul and Trimble (1970) speculated that the inferiority of taped training is a function of the loss of response-contingent progression through the training procedure. This hypothesis has yet to be adequately tested, although Riddick and Meyer (1973) found that taped training with response-contingent (gross motor movement) feedback was as effective as live training. Generalization of that finding to the clinical setting is limited because of the study's non-clinical situational context and subject sample, as well as the administration of only one training session. Further research, therefore, might examine the basis for the superiority of live over taped modes of training. Hypotheses of interest might include presence/absence of a therapist as well as response-contingent versus non-response-contingent progression through the procedure.

The comparison of self-relaxation and EMG biofeedback indicated that both types of training led to significant reductions in physiological arousal and subjective tension. Interestingly, SR was differentially more effective than BF in accomplishing heart rate reductions during training. This may reflect the participants' general ability to decrease heart rate in the absence of elaborate training procedures (Ray & Lamb, 1974). These data indicate that a rather elaborate biofeedback procedure involving complex and expensive equipment is no more, and possibly less, effective in producing relaxation during training than a simple self-relaxation procedure. Our BF procedure was individualized for each client by virtue of the shaping procedure used. It has been shown that the auditory feedback/eyes-closed mode of training is superior to other modes of feedback (Alexander, French, & Goodman, 1975). Yet self-relaxation was superior to feedback on one autonomic dependent variable in the present study. Previous

research on normal subjects has indicated that EMG biofeedback is superior to simple instructions to relax when EMG is the primary dependent variable (Coursey, 1975; Haynes, Moseley, & McGowan, 1975). The present results suggest that the addition of a therapeutic rationale and daily practice of relaxation skills to self-relaxation instructions yields a procedure that is at least as effective as EMG biofeedback for producing in-session relaxation in chronically anxious subjects. In the present study the superiority of self-relaxation for reducing heart rate, when considered with earlier cautions about the effectiveness of EMG biofeedback as a general relaxation training technique (Alexander, 1975; Shedivy & Kleinman, 1977), suggests that such feedback should not be used clinically to attempt to provide a general state of relaxation. For future research on how to best accomplish in-session relaxation, rather than pursue costly and questionably effective feedback as an alternative to extended progressive relaxation training in the clinical setting, it would seem preferable to investigate the potential benefits of self-relaxation training.

Posttraining Effects

This is the first investigation to report data regarding client control of tonic physiological arousal and subjective tension after relaxation training procedures have been completed. The comparison of all four training procedures on posttraining control indicated that all groups were able to significantly reduce muscle and subjective tension when asked to do so, with no differential change across groups. With respiration rate, the system under direct voluntary control (Paul, 1969), live progressive relaxation and self-relaxation training were both superior to taped progressive relaxation and biofeedback training. Only live extended progressive relaxation training led to client control over frequency of electrodermal response, a system enervated solely by the sympathetic nervous system (Sternbach, 1966). Differential patterning of control (reductions in somatic and subjective response, increases in sympathetic response) was true for all groups but live progressive relaxation. This could indicate that the other groups may have

been "actively" trying to relax. They were successful for the somatic (muscle tension, respiration rate) and subjective systems but not for the electrodermal measure of sympathetic arousal.

Considering the generally accepted operational definition of anxiety as subjective distress accompanied by sympathetic arousal (Paul, 1969), the fact that live progressive relaxation training alone led to reductions in sympathetic arousal seems particularly important. Since these clients were selected for their reported difficulties with anxiety and tension in their everyday lives, live extended progressive relaxation training would appear to be the clinical treatment of choice when client control over tonic sympathetic arousal is one of the clinical objectives.

In a separate investigation in the same laboratory, live abbreviated relaxation training was superior to self-relaxation training in analogue clinical subjects on subjective response to phobic stimuli (Beiman, Green, Webster, Rosmarin, Holliday, & Graham, Note 1). Generalization of the present investigation's results to other relaxation procedures, for example, those involving direct induction, should not be assumed and awaits empirical verification. In the interim, considering the present data as well as those of Beiman et al. (Note 1), it seems appropriate to recommend live extended progressive relaxation training as used here for the treatment of chronic or pervasive anxiety/tension.

The use of this treatment in behavioral medicine appears to have considerable promise. Since stress (involving prolonged adreno-sympathetic arousal) has been shown to produce tissue damage and contribute to an extensive and varied list of medical disorders (Selye, 1976), the use of extended relaxation training in the psychological treatment of such stress-related disorders seems particularly appropriate and has been recommended previously (Schwartz & Shapiro, 1973). The present investigation provides empirical support for such a recommendation, with the qualifications regarding specific type of training discussed above kept in mind. Extended live progressive relaxation training presented as a self-control coping skill for maladaptive tension and anxiety has also proven successful in the

nonpharmacological treatment of essential hypertension (Graham, Beiman, & Ciminero, 1977; Beiman, Graham, & Ciminero, Note 2). This and other stress-related medical disorders are potentially good targets for future outcome research on the effects of live extended progressive relaxation training.

Reference Notes

1. Beiman, I., Green, K., Webster, J., Rosmarin, D., Holliday, P., & Graham, L. A comparison of abbreviated progressive and self-relaxation training on measures of tonic physiological arousal and phasic response to fearful stimuli. Manuscript submitted for publication, 1977.
2. Beiman, I., Graham, L., & Ciminero, A. Self-control progressive relaxation training for essential hypertension: Therapeutic effects in the natural environment. Manuscript submitted for publication, 1977.

References

- Alexander, A. B. An experimental test of assumptions relating to the use of electromyographic biofeedback as a general relaxation training technique. *Psychophysiology*, 1975, 12, 656-662.
- Alexander, A. B., French, C. A., & Goodman, N. J. A comparison of auditory and visual feedback in biofeedback-assisted muscular relaxation training. *Psychophysiology*, 1975, 12, 119-123.
- Beary, J. F., Benson, H., & Klemchuk, H. P. A simple psychophysiological technique which elicits the hypometabolic changes of the relaxation response. *Psychosomatic Medicine*, 1974, 36, 115-120.
- Benson, H. *The relaxation response*. New York: Morrow, 1976.
- Benson, H., Greenwood, M. M., & Klemchuk, H. The relaxation response: Psychophysiological aspects and clinical applications. *International Journal of Psychiatry in Medicine*, 1975, 6, 87-98.
- Bernstein, D. A., & Borkovec, T. D. *Progressive relaxation training: A manual for the helping professions*. Champaign, Ill.: Research Press, 1973.
- Coursey, R. D. Electromyograph feedback as a relaxation technique. *Journal of Consulting and Clinical Psychology*, 1975, 43, 825-834.
- Cox, D. J., Freundlich, A., & Meyer, R. G. Differential effectiveness of electromyograph feedback, verbal relaxation instructions, and medication placebo with tension headaches. *Journal of Consulting and Clinical Psychology*, 1975, 43, 892-898.
- Duncan, D. B. Multiple-range and multiple F tests. *Biometrics*, 1955, 11, 1-42.
- Edelberg, R. Electrical properties of the skin. In C. Brown (Ed.), *Methods in psychophysiology*. Baltimore, Md.: Williams & Wilkins, 1967.
- Graham, L., Beiman, I., & Ciminero, A. The generality of the therapeutic effects of progressive relaxation training for essential hypertension. *Journal of Behavior Therapy and Experimental Psychiatry*, 1977, 8, 161-164.
- Haynes, S. N., Moseley, D., & McGowan, W. T. Relaxation training and biofeedback in the reduction of frontalis muscle tension. *Psychophysiology*, 1975, 12, 547-552.
- Husek, T. R., & Alexander, S. The effectiveness of the Anxiety Differential in examination situations. *Educational and Psychological Measurement*, 1963, 23, 309-318.
- Israel, E., & Beiman, I. Live versus recorded relaxation training: A controlled investigation. *Behavior Therapy*, 1977, 8, 251-254.
- Kirk, R. E. *Experimental design: Procedures for the behavioral sciences*. Belmont, Calif.: Brooks/Cole, 1968.
- Lykken, D., & Venables, P. Direct measurement of a skin conductance: A proposal for clarification. *Psychophysiology*, 1971, 8, 656-672.
- Mathews, A. M. Psychophysiological approaches to the investigation of desensitization and related procedures. *Psychological Bulletin*, 1971, 76, 73-91.
- Paul, G. L. Physiological effects of relaxation training and hypnotic suggestion. *Journal of Abnormal Psychology*, 1969, 74, 425-437.
- Paul, G. L., & Trimble, R. W. Recorded vs. "live" relaxation training and hypnotic suggestion: Comparative effectiveness for reducing physiological arousal and inhibiting stress response. *Behavior Therapy*, 1970, 1, 285-302.
- Ray, W. J., & Lamb, S. B. Locus of control and the voluntary control of heart rate. *Psychosomatic Medicine*, 1974, 36, 180-182.
- Reinking, R. H., & Kohl, M. L. Effects of various forms of relaxation training on physiological and self-report measures of relaxation. *Journal of Consulting and Clinical Psychology*, 1975, 43, 595-600.
- Riddick, C., & Meyer, R. C. The efficacy of automated relaxation training with response-contingent feedback. *Behavior Therapy*, 1973, 4, 331-337.
- Russell, R., Sipich, J., & Knipe, J. Progressive relaxation training: A procedural note. *Behavior Therapy*, 1976, 1, 566-568.
- Schwartz, G., & Shapiro, D. Biofeedback and essential hypertension: Current findings and theoretical concerns. *Seminars in Psychiatry*, 1973, 5, 493-503.
- Selye, H. *Stress in health and disease*. Woburn, Mass.: Butterworth, 1976.
- Shedivy, D. I., & Kleinman, K. M. Lack of correlation between frontalis EMG and either neck EMR or verbal ratings of tension. *Psychophysiology*, 1977, 14, 182-186.
- Spielberger, C., Gorsuch, R., & Lushene, R. *Manual for the State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1968.
- Sternbach, R. *Principles of psychophysiology*. New York: Academic Press, 1966.
- Venables, P. H., & Martin, F. *Manual of psychophysiological methods*. New York: American Elsevier, 1967.
- Zuckerman, M., & Lubin, B. *Manual for the Multiple Affect Adjective Check List*. San Diego, Calif.: Educational and Industrial Testing Service, 1965.

Correlational and Factor Analysis of the Peabody Individual Achievement Test and the WISC-R

Richard L. Wikoff
University of Nebraska at Omaha

Subtest scores from the Peabody Individual Achievement Test (PIAT) and the Wechsler Intelligence Scale for Children-Revised (WISC-R) for 180 children, ages 6-17, were factor analyzed to determine the number and kinds of factors measured by the PIAT. Two factors were found when the PIAT was factored alone. Reading Recognition, Reading Comprehension, and Spelling loaded highly on a Word Recognition factor, whereas Mathematics and General Information had moderate to high loadings on a School-related Knowledge factor. When the PIAT was factored with the WISC-R subtests, a Word Recognition factor was found in addition to the three factors usually reported for the WISC-R. The School-related Knowledge factor of the PIAT was subsumed by the other factors. General Information loaded highly on the Verbal Comprehension factor, and Mathematics loaded highly on the Freedom from Distractibility factor. Implications for the interpretation of the PIAT are discussed.

The Peabody Individual Achievement Test (PIAT; Dunn & Markwardt, 1970) has gained wide acceptance and use during the past several years. It contains five subtests that are separately scored and have separate norms for both age and grade. The materials for the subtests are presented either verbally or by presenting a page of material for the subject to read. The subject responds by choosing from among four choices, which are visually presented. Raw scores are determined by the number of correct responses and can be converted to either percentile ranks or standard scores with a mean of 100 and a standard deviation of 15. In addition, there are normative tables for the total test score.

The Mathematics subtest was designed as a measure of the ability to apply mathematical concepts to the solution of problems, as opposed to a measure of computational skills or quantitative aspects of concept formation. The objective of the Reading Recognition subtest "is to measure skills in translating sequences of printed alphabetic symbols which form words into speech sounds that can be

understood by others as words" (Dunn & Markwardt, 1970, p. 19). The Reading Comprehension subtest is presented by its authors as a test of the ability to derive meaning from printed words. The Spelling subtest attempts to measure the ability to recognize correct spelling as opposed to the ability to recall the correct spelling of words. The General Information subtest is presented as a measure of general encyclopedic knowledge.

The interpretation of these five subtest scores depends on the extent to which they are measuring different factors. It is quite possible that only one factor, for example, general intelligence, is present. In this case, the best interpretations would be made from the total test score, since this score would be the most reliable. On the other hand, if the subtests are measuring five different factors, interpretations based on each separate subtest would be appropriate.

One of the most widely used intelligence tests is the Wechsler Intelligence Scale for Children-Revised (WISC-R, Wechsler, 1974). The factor structure for this test was reported by Kaufman (1975), who found three meaningful factors—(a) Verbal Comprehension, (b) Perceptual Organization, and (c) Freedom from Distractibility.

Requests for reprints should be sent to Richard L. Wikoff, Department of Psychology, University of Nebraska, Box 688, Omaha, Nebraska 68101.

The purpose of this study was to determine the number of factors being measured by the PIAT subtests and the extent to which each subtest measured the factors found. A secondary purpose was to investigate the relationship of the PIAT to the WISC-R to determine if it measures some aspect of intelligence or measures factors not found in the WISC-R.

Method

Subjects

The subjects of this study were children who were referred to me because of learning problems. Each subject's classroom achievement was below either his or her parents' or the school's expectations. Nearly all were described as hyperactive, easily distracted, having a short attention span, not completing assignments, and so on. The children ranged in age from 6 years to 17 years, with a median age of 9.87 and a mean age of 10.1. There were 123 males and 57 females. No minority children were included. All socioeconomic groups were represented, but most were from the middle class.

Procedure

Each subject was administered the WISC-R and the PIAT. Only the protocols for those subjects who had scores on 10 WISC-R subtests (excluding Digit Span and Mazes) and all five PIAT subtests were used. There were 180 subjects who met this criterion.

A principal factors solution was used to factor analyze the standard scores for the PIAT subtests. Communalities were determined by iteration, and the resulting solution was rotated using the varimax method. Rotations were performed for one, two, and three factors to determine the most appropriate structure (Gorsuch, 1974).

A second factor analysis was performed that included standard scores for 10 subtests of the WISC-R. This was done to determine if the PIAT was actually measuring achievement or if it was measuring the factors related to intelligence as found in the WISC-R (Kaufman, 1975). Varimax rotations were performed with three, four, and five factors. The four-factor solution was the most meaningful.

Results

Two factors were retained for the PIAT subtests. Reading Recognition, Reading Comprehension, and Spelling loaded highly on Factor 1. Since all three subtests require recognition of words, this factor was named *Word Recognition*. General Information and Mathematics had the highest loadings on Factor 2. Both of these subtests require a knowledge of facts acquired in school;

Table 1

Factor Structure of the Peabody Individual Achievement Test Subtests for Learning-Problem Children

Subtest	Word Recognition	School-related Knowledge	h^2
Mathematics	.33	.66	.54
Reading Recognition	.91	.37	.97
Reading Comprehension	.73	.50	.78
Spelling	.70	.40	.65
General Information	.36	.71	.63
Eigenvalue	2.09	1.49	3.58
% of h^2	58.38	41.62	100.00
% of total variance	41.80	29.80	71.60

Note. $N = 180$.

therefore, this factor was named *School-related Knowledge*. The factor loadings are presented in Table 1.

When the PIAT and WISC-R subtests were factor analyzed together, four factors were retained. The structure is presented in Table 2 along with means and standard deviations for each of the variables. The first factor had high to very high loadings for Reading Recognition, Reading Comprehension, and Spelling. This factor was clearly the Word Recognition factor from the PIAT and was obviously measuring a factor that is only barely tapped by the WISC-R subtests. The second factor had the highest loadings for Vocabulary, Comprehension, Similarities, and Information. It was labeled *Verbal Comprehension*, since it corresponds to the first factor of the WISC-R (Kaufman, 1975). The PIAT subtests had only low loadings on this factor, with the exception of General Information, which loaded highly. The third factor corresponds to the Perceptual Organization factor of the WISC-R (Kaufman, 1975). Its highest loadings were for Object Assembly, Block Design, Picture Completion, and Picture Arrangement. Mathematics and Arithmetic had the highest loadings on Factor 4. This factor appears to be similar to the Freedom from Distractibility factor of the WISC-R (Kaufman, 1975). The loadings were smaller than those reported by

Table 2

Means, Standard Deviations, and Factor Structure for the WISC-R and PIAT Subtests for Learning-Problem Children

Subtest	<i>M</i>	<i>SD</i>	Factor				<i>N</i> ²
			1: Word Recogni- tion	2: Verbal Compre- hension	3: Perceptual Organiza- tion	4: Freedom from Distracti- bility	
Information	9.44	2.87	.34	.66	.15	.22	.62
Similarities	10.17	2.80	.34	.64	.33	.07	.64
Arithmetic	9.03	2.78	.37	.41	.28	.57	.71
Vocabulary	10.28	3.21	.29	.81	.22	.10	.80
Comprehension	10.42	2.89	.08	.66	.34	.30	.65
Picture Completion	10.71	2.72	.09	.34	.60	.00	.48
Picture Arrangement	10.43	2.96	.08	.35	.50	.16	.40
Block Design	10.09	3.18	.17	.15	.77	.08	.65
Object Assembly	10.74	3.09	.03	.09	.78	.13	.63
Coding	8.43	3.23	.18	.09	.35	.11	.18
Mathematics	96.69	12.63	.38	.34	.22	.58	.64
Reading Recognition	96.83	12.87	.91	.29	.19	.09	.96
Reading Comprehension	96.54	13.45	.75	.35	.23	.19	.77
Spelling	93.89	14.00	.77	.21	.07	.21	.69
General Information	99.94	12.45	.39	.72	.20	.22	.76
Eigenvalue			2.82	3.72	2.49	1.01	9.59
% of <i>h</i> ²			29.41	34.10	25.96	10.53	100.00
% of total variance			18.80	21.80	16.60	6.73	63.93

Note. WISC-R = Wechsler Intelligence Scale for Children-Revised; PIAT = Peabody Individual Achievement Test; *N* = 180.

Kaufman because some of the variance was rotated to other factors.

The Reading Recognition, Reading Comprehension, and Spelling subtests were all highly correlated with the Word Recognition factor. These subtests had only low correlations with the other factors, except that Reading Comprehension correlated .35 with the Verbal Comprehension factor.

The General Information subtest correlated .72 with Verbal Comprehension and .39 with Word Recognition. Correlations with the other factors were low.

The Mathematics subtest correlated the highest with the Freedom from Distractibility factor but had loadings of .38 and .34, respectively, for Word Recognition and Verbal Comprehension.

Another way of looking at the relationship of the PIAT to the WISC-R is by the correlation of the PIAT subtests with the Verbal, Performance, and Full Scale IQs of the WISC-R. These correlations are presented in Table 3.

Discussion

The structure of the PIAT subtests has not been reported previously. The results of this study indicate that there are actually only two factors rather than the five implied by the subtest organization. Factor 1 was a Word Recognition factor that included the two reading subtests and Spelling. Factor 2 was a School-related Knowledge factor that included Mathematics and General Information.

When the PIAT subtests were analyzed with the WISC-R subtests, the Word Recognition factor remained as a factor distinct from those previously reported for the WISC-R (Kaufman, 1975). The second factor was subsumed by the WISC-R factors. General Information had its highest loading on the Verbal Comprehension factor, and Mathematics loaded along with the WISC-R Arithmetic subtest on the Freedom from Distractibility factor.

The results of this analysis are not conclusive, but they suggest the possibility that the

Table 3
Correlations of PIAT Subtests with
WISC-R Verbal, Performance, and
Full Scale IQs

Subtest	Verbal	Performance	Full Scale
Mathematics	.57	.39	.57
Reading Recognition	.58	.39	.56
Reading Comprehension	.64	.43	.60
Spelling	.50	.28	.43
General Information	.78	.44	.71

Note. PIAT = Peabody Individual Achievement Test; WISC = Wechsler Intelligence Scale for Children-Revised. $N = 180$.

Freedom from Distractibility factor may actually be a numerical factor. This follows from two lines of reasoning. First, in Kaufman's (1975) analysis of the WISC-R, Arithmetic, Digit Span, and Coding (all tasks involving numbers) had moderate loadings on the Freedom from Distractibility factor. Second, the PIAT Mathematics subtest and the WISC-R Arithmetic subtest have similar loadings on this factor. However, the PIAT Mathematics subtest appears to minimize the need for concentration by presenting material visually, which the subject can refer to while solving the problem. The WISC-R Arithmetic subtest requires the subject to retain information presented orally. This problem might be investigated further by factor analyzing these tests with numerical tests whose structure is known.

Since the two reading subtests and the Spelling subtest of the PIAT are measuring the same factor, they should be interpreted together. Remedial treatment designed to

increase the scores on any one of these subtests should improve the scores on the other two as well. To increase reliability, it is suggested that these three subtests be combined and new norms presented for the new variable, Word Recognition. The other two subtests, Mathematics and General Information, can be appropriately interpreted as separate subtests. The strong relationship to Verbal Comprehension should be kept in mind when interpreting scores from the latter test. Further study is recommended to determine if persons taking this subtest might know the information sought but are unable to understand the language.

Finally, these results support the use of the PIAT as a separate test in a battery containing the WISC-R. There is at least one factor that is different from those measured by the WISC-R that provides additional information of value in developing a treatment program for helping children with learning problems. The Mathematics subtest appears to provide supplementary information, also. However, the General Information subtest correlates so highly with Verbal Comprehension that it would not be necessary to include this subtest.

References

- Dunn, L. M., & Markwardt, F. C., Jr. *Peabody Individual Achievement Test Manual*. Circle Pines, Minn.: American Guidance Service, 1970.
- Gorsuch, R. L. *Factor analysis*. Philadelphia, Pa.: Saunders, 1974.
- Kaufman, A. S. Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology*, 1975, 43, 135-147.
- Wechsler, D. *Manual for the Wechsler Intelligence Scale for Children-Revised*. New York: Psychological Corporation, 1974.

Received September 20, 1977 ■

Comments

Social Psychological Concepts Applied to Clinical Processes: On the Need for Precision

Ben Harris
Radford College

John H. Harvey
Vanderbilt University

This comment analyzes and critiques an attempted application of cognitive dissonance and reactance concepts to a therapy analogue. An experiment by R. M. Gordon is discussed. It is shown that key prerequisites for dissonance arousal (choice and responsibility) and for reactance induction (initial free choice) were not present. Also, self-selection of volunteers for treatment is shown to be a likely, unreported factor. Original results are reinterpreted to suggest that they were due to the frustration of some subjects' expectations rather than to arousal of reactance. A need is cited for well-informed application of experimental social psychological theories.

Recently, clinical psychologists (e.g., Brehm, 1976; Goldstein, Heller, & Sechrest, 1966) have offered theoretical analyses concerning the role of two social psychological processes (cognitive dissonance and psychological reactance) in mediating patient attitudes toward clinical treatment. Although this work has not yet had a profound impact on the practice of clinical psychology, it reflects clinicians' increased attention to social and intrapersonal processes that may affect psychotherapy theory and practice. Consistent with this emphasis, Gordon (1976) has attempted to apply cognitive dissonance and reactance theory to the important clinical variables of incentive to engage in therapy and choice of treatment. Unfortunately, he has primarily succeeded in illustrating the difficulties involved in such an endeavor. This comment criticizes the internal and external validity of Gordon's study, offers an alternative explanation for his findings, and attempts to clarify some of the conceptual issues that are raised.

Subjects

In Gordon's study, volunteers and nonvolunteers were recruited for a session of relaxation

The authors would like to thank Robert Falk for his assistance in the preparation of this comment. Thanks are also due to Robert Gordon for his gracious cooperation in answering inquiries by the first author.

Requests for reprints should be sent to Ben Harris, Department of Psychology, Radford College, Radford, Virginia 24142.

training and were given a choice between two types of training (choice conditions) or were assigned to a treatment type chosen by another subject (no-choice conditions). A major challenge to the validity of Gordon's study is posed by his manner of operationalizing the concept of volunteering for treatment. In this experiment, volunteer and nonvolunteer subjects were students from a single undergraduate psychology class. On a scale of self-reported relaxation, the mean pretreatment score for all subjects was 6.8 (out of a maximum 10), suggesting that these subjects were not comparable to a clinical population. More importantly, both "volunteer" and "nonvolunteer" groups received the same amount of course credit for their participation, both groups were apparently recruited during class hours in the same week, and many volunteers and nonvolunteers received the experimental treatment in pairs (Gordon, Note 1). Thus, the act of volunteering for treatment was one with seemingly no subsequent discriminative effect on any of the subjects: The subjects made no differential sacrifice after volunteering, they received no differential incentive, and they were given no preferential choice of treatment. Thus, any experimental effect found for the factor of volunteering/nonvolunteering would be most likely due to the self-selection of subjects into volunteer and nonvolunteer groups rather than due to an actual treatment effect (i.e., dissonance arousal). In regard to this self-selection hypothesis, the male-to-female ratio for volunteer subjects was approximately 1:1, whereas it was 3:1 for nonvolunteer subjects. Also, there

was a 10% difference in the self-reported anxiety of volunteers and nonvolunteers. These differences between subjects in the volunteer and nonvolunteer conditions are consistent with Rosenthal and Rosnow's (1975, pp. 13-25, 56-59) review of self-selection effects in experimental and clinical research.

Subsequent to the selection of volunteers and nonvolunteers for relaxation training, all subjects in the Gordon study (Gordon, 1976) were exposed (in randomly constructed pairs) to the principal experimental manipulation—the "responsibility for choice" of treatment modality. This manipulation was accomplished by having one member of each pair of subjects choose between two types of relaxation training; then, both subjects were exposed to the chosen treatment. The major theoretical reason for this manipulation was that feelings of personal responsibility represent one of many necessary conditions for producing cognitive dissonance (Wicklund & Brehm, 1976, pp. 51-71). Unfortunately, the study examined here does not directly manipulate responsibility; it simply gives a choice between two therapy treatments to half of the subjects and no choice to the others. Even though perceived choice may have been varied (cf. Harvey & Harris, 1975) and may have been related to responsibility (Harvey, Harris, & Barnes, 1975), Gordon (Note 2) did not use a manipulation check to assess his subjects' actual feelings of either responsibility or choice. As a result, at least one important prerequisite for dissonance arousal may have been absent.

In evaluating Gordon's application of cognitive dissonance theory, an even more important methodological question concerns the nature of the treatment offered to subjects. In introducing the choice manipulation, Gordon argued that making an irrevocable choice between two evenly matched alternatives will produce cognitive dissonance, since the chooser's subsequent behavior is more partial to the chosen alternative than its relative attractiveness would allow. Thus, in an experimental situation similar to the one in question, Gordon would predict subjects' engaging in dissonance-reducing behavior such as selective perception of choice-related information (Ehrlich, Guttman, Schonbach, & Mills, 1957) or heightened evaluation of the selected treatment modality. Unfortunately, Gordon's methodology does not allow a test of this hypothesis, since his subjects did not choose between *evenly matched alternatives* (it was found that one treatment was preferred to the other). Thus, a second impor-

tant prerequisite for dissonance arousal was absent.

In addition to cognitive dissonance theory, Gordon invoked J. W. Brehm's (1966) theory of psychological reactance to explain his experimental results. To do this, he hypothesized that low-responsibility (no-choice) subjects will experience reactance because of their limited control over the treatment they receive. As with the dissonance-based explanation, Gordon's assumption about the involvement of a reactance process cannot be tested using his experimental method. Theoretically, reactance is induced by eliminating one or more of a person's existing free (possible) behaviors (J. W. Brehm, 1966, p. 4). In Gordon's study, the "low-responsibility" subjects were not first given choices and then restricted in their options; they were simply given a treatment selected by another subject. According to reactance theory, one would not expect this type of variation to result in either attempts to restore choice or in devaluation of the imposed treatment. Also, as with dissonance, Gordon provided no data to demonstrate *directly* the operation of a reactance process in his study.

If cognitive dissonance were aroused by Gordon's manipulation of subjects' choice, one would expect a significant main effect of this factor (high choice vs. no choice) for subjects' attitudes toward treatment. Gordon's failure to find a main effect of choice/no choice for the measure of subjects' attitudes would seem to be consistent with the above critique of the study's dissonance-related methodology.

Gordon invoked reactance theory to explain his findings for the measure of subjects' self-rated change in relaxation following treatment. He found that only volunteers with no choice of treatment failed to show a significant (post-treatment) increase in relaxation. As the above examination of the method suggests, this result is most likely *not* due to psychological reactance being aroused in the no-choice volunteers. A more likely explanation would center on the frustration experienced by the no-choice volunteers. A review of Gordon's procedure shows that volunteer subjects both expressed interest in the treatment earlier than nonvolunteers and were somewhat less relaxed ($M = 6.3$) than nonvolunteers ($M = 7.4$) before the treatment. Accordingly, they may have both desired and expected preferential treatment. For those volunteers in the no-choice condition, however, their treatment was picked by another subject who then received his or her preferred treatment simultaneously and in the same room as the no-

choice volunteer. Thus, the effect of this manipulation (less self-reported change in relaxation) can be more parsimoniously explained by attending to the relative powerlessness and frustrated expectations of the no-choice volunteer subjects.

We are not arguing here that cognitive dissonance and psychological reactance are irrelevant to theorizing about clinical variables such as attitude toward treatment and self-perceived therapeutic change. However, the elaborate previous work on these variables and theories in the area of social psychology necessitates considerable caution when making applications to clinical concerns. It is hoped that the continuing interest of clinicians in experimental social psychology (e.g., S. Brehm, 1976) will promote *well-informed* attempts to apply social psychological ideas to clinical problems and processes.

Reference Notes

1. Gordon, R. M. Personal communication, February 1977.
2. Gordon, R. M. Personal communication, December 1976.

References

Brehm, J. W. *A theory of psychological reactance*. New York: Academic Press, 1966.

- Brehm, S. *The application of social psychology to clinical practice*. New York: Halstead Press, 1976.
- Ehrlich, D., Guttman, I., Schonbach, P., & Mills, J. Postdecision exposure to relevant information. *Journal of Abnormal and Social Psychology*, 1957, 54, 98-102.
- Goldstein, A. P., Heller, K., & Sechrest, L. B. *Psychotherapy and the psychology of behavior change*. New York: Wiley, 1966.
- Gordon, R. M. Effects of volunteering and responsibility on the perceived value and effectiveness of a clinical treatment. *Journal of Consulting and Clinical Psychology*, 1976, 44, 799-801.
- Harvey, J. H., & Harris, B. Determinants of perceived choice and the relationship between perceived choice and expectancy about feelings of internal control. *Journal of Personality and Social Psychology*, 1975, 31, 101-106.
- Harvey, J. H., Harris, B., & Barnes, R. D. Actor-observer differences in the perceptions of responsibility and freedom. *Journal of Personality and Social Psychology*, 1975, 32, 22-28.
- Rosenthal, R., & Rosnow, R. L. *The volunteer subject*. New York: Wiley, 1975.
- Wicklund, R. A., & Brehm, J. W. *Perspectives on cognitive dissonance*. Hillsdale, N.J.: Erlbaum, 1976.

Received February 8, 1977 ■

Imprecision or Dissonance? A Reply to Harris and Harvey

Robert M. Gordon
Allentown, Pennsylvania

Harris and Harvey's criticism of Gordon's research as imprecise is based on their misunderstanding of my procedures and making much ado about initial nonsignificant differences between the volunteers and the nonvolunteers. Using this to explain away the volunteer factor, they claim that the prerequisites of dissonance and reactance were not present. I suggest that the real issue is the efficacy of volunteering as an independent variable, not imprecision.

Harris and Harvey (1978) expressed a fear that clinical researchers will apply social psychological theories without due regard or understanding. As a social clinician whose predoctoral specialty was in social psychology, I don't think the problem is with sloppy cross-fertilization. Harvey and Harris (1975; Harvey, Harris, & Barnes, 1975), who have done interesting work with the choice and responsibility variables, must feel that my study (Gordon, 1976), which deals in part with the same variables, cannot also be true. Unfortunately, they appear to reconcile our supposed differences by making much ado about nonsignificant differences and misinterpreting my procedures. The only real disagreement, as I see it, is the efficacy of the volunteering variable—a theoretical issue that they avoid while focusing on perceived impreciseness.

Harris and Harvey infer that my study (Gordon, 1976) has no external validity, since the subjects used were college students rather than a clinical population—perhaps believing that dissonance and reactance are limited to college students. As evidence, they stated that the initial pretreatment mean for relaxation was 6.8. As a casual demonstration, I asked 14 of my private clients with anxiety problems the same question as was used in the Gordon study: "How relaxed are you right now?" (1 = "not at all" and 10 = "completely"). The results were very similar ($M = 6.9$, $SD = 1.0$), not surprising to anyone who makes a distinction between state and trait variables. The two are not necessarily correlated. Harris and Harvey do not feel that volunteering is an indication of motivation involving

choice, responsibility, or interest. And since they don't, they claim that the prerequisites of dissonance and reactance were not present. They claim instead that the volunteer effects are just a methodological fluke, due to initial group differences in sex and anxiety. However, sex was not found to be significantly correlated with any of the dependent measures; and the 10% difference in initial anxiety between the volunteers and nonvolunteers was also nonsignificant, $F(1, 26) = 1.84$. They also did not attempt to explain why the volunteers continued to volunteer *after* the relaxation experiment, when again there were no volunteer main-effect differences.

Harris and Harvey (1978) misinterpret the procedure as it "simply gives a choice between two therapy treatments to half of the subjects and no choice to the others" (p. 327). They conclude that both dissonance and reactance could not have existed. In reducing the actual procedure to a simple choice—no choice situation, they have eliminated the entire relationship process, which was the reason for running the subjects in pairs—to accentuate the differential feelings of responsibility and choice. Both subjects were told about two very different forms of relaxation training that were available to them and that they could have only one of them. The experimenter asked one of the subjects which form he/she would prefer and then played that tape without considering the other subject. The subject who had chosen the tape chose it for the other subject as well. The yoked other was not only unfairly denied the right to choose, but he/she also had to hear a tape chosen by the other subject. As predicted, the inferred feelings of dissonance in the former and reactance in the latter occurred for only those students who volunteered for treatment. The dissonance (high-responsibility-and-choice) group valued the treat-

Requests for reprints should be sent to Robert M. Gordon, 1251 South Cedar Crest Boulevard, Allentown, Pennsylvania 18103.

ment highly and claimed that it was very successful, whereas the reactance group devalued the treatment and was the only group of four who found the treatment to be unsuccessful. The nonvolunteers, who were not particularly interested in the treatment to begin with, were unaffected by the subsequent experimental manipulation.

Harris and Harvey claim that the subjects did not choose between evenly matched alternatives. They must have misunderstood my comment: "Surprisingly, many of the subjects had a definite preference" (Gordon, 1976, p. 800), taking it to mean that there was a *consistently* clear preference for one alternative over the other. This was not the case. I was surprised because there are no such forms of relaxation as "neuroglandular" and "cardiovascular." Yet people had definite preferences. Harris and Harvey contend that the processes of dissonance and reactance were not directly measured. This is true, but as far as I know, no theoretical construct is ever directly measured—it is only inferred.

Harris and Harvey claim that since there was no main effect of choice, they are correct in their argument. This circular reasoning assumes that the volunteering effects are not real. Harris and Harvey's alternative hypothesis erroneously assumes (a) that the volunteers expressed an earlier interest than the nonvolunteers; however, the volunteers and nonvolunteers expressed their differential attitudes at the same time. Harris and Harvey may be confusing the later manipulation to get the nonvolunteers into treatment with their initial attitudes. Also, the subjects' initial attitudes remained unchanged even after the treatment. They never became "volunteers." (b) The volunteers were less relaxed. Again, there were no significant volunteer pretreatment or posttreatment main effects in anxiety. (c) The volunteers expected preferential treatment. This, I suppose, is based on the above two assumptions. Treatment was not based on need or on a first-come, first-served basis.

Conclusions

Theoretically I would agree with Harris and Harvey's criticisms if their interpretations were

indeed true. The problem is certainly not in the improper understanding of social psychological theories. In attempting to reconcile perceived mutually exclusive findings, they have reinterpreted my study (Gordon, 1976) to seem imprecise. This is not an unusual form of dissonance reduction for researchers; Witness the plethora of straw-man controversies.

The only theoretical difference we may have is whether volunteering represents a methodological problem or an independent variable in its own right (Rosenthal & Rosnow, 1975). The volunteer issue has been a controversial one for some time. Perhaps part of the reason it has been so unwelcome is that it is so humbling. It infers that our procedures are only effective if people are willing to let them be effective. It infers that their initial motivations, feelings of personal responsibility, and resistances are more powerful than the treatment per se. The clinician, who has long observed this in therapy, cannot explain this away as self-selection or randomize or factor it away. The clinician must deal with these dynamics if the treatment is going to work. Volunteering is not a mistake—it is choice and responsibility.

References

- Gordon, R. M. Effects of volunteering and responsibility on the perceived value and effectiveness of a clinical treatment. *Journal of Consulting and Clinical Psychology*, 1976, 44, 799-801.
- Harvey, J. H., & Harris, B. Determinants of perceived choice and the relationship between perceived choice and expectancy about feelings of internal control. *Journal of Personality and Social Psychology*, 1975, 31, 101-106.
- Harvey, J. H., Harris, B., & Barnes, R. D. Actor-observer differences in the perceptions of responsibility and freedom. *Journal of Personality and Social Psychology*, 1975, 32, 22-28.
- Harris, B., & Harvey, J. H. Social psychological concepts applied to clinical processes: On the need for precision. *Journal of Consulting and Clinical Psychology*, 1978, 46, 326-328.
- Rosenthal, R., & Rosnow, R. L. *The volunteer subject*. New York: Wiley, 1975.

Received April 27, 1977

Selecting a Short Form of the MMPI: Addendum to Faschingbauer

Norman G. Poythress, Jr.

Center for Forensic Psychiatry, Ann Arbor, Michigan

Faschingbauer offered guidelines to clinicians in the selection of Minnesota Multiphasic Personality Inventory (MMPI) short forms as substitutes for the full MMPI. This comment offers an addendum to Faschingbauer in the form of a review of empirical studies of the clinical validity of MMPI short forms and a discussion of the MMPI-168, which was not considered in Faschingbauer's earlier article. For diagnostic and interpretive accuracy, the empirical evidence to date seems to favor two short forms—the Faschingbauer Abbreviated MMPI and the MMPI-168—over the other available short forms.

In a recent review Faschingbauer (1976) discussed numerous clinical considerations one should make in selecting a short-form Minnesota Multiphasic Personality Inventory (MMPI) for clinical use in place of the full MMPI. In his article, Faschingbauer offered a critical analysis of four recently developed short forms—the Mini-Mult (Kincannon, 1968), the Hugo (Hugo, 1971), the Midi-Mult (Dean, 1972), and the Faschingbauer Abbreviated MMPI (FAM; Faschingbauer, 1974)—and addressed a variety of issues including (a) correlations of short-form scales with the full MMPI, (b) differences in item content, and (c) concordance rates between short forms and the full MMPI for Welsh code types.

Faschingbauer's review offered excellent guidelines for the clinician who may consider using an MMPI short form. If shortcomings are to be noted in his review, they would have to be that (a) he failed to include the MMPI-168 (Overall & Gomez-Mont, 1974) in his discussion, and (b) he excluded from discussion studies on the clinical (as opposed to statistical) validity of the various short forms. This comment offers an addendum to Faschingbauer (1976) and includes a discussion of the MMPI-168 and a review of the clinical validity studies of the MMPI short forms.

MMPI-168

The development and description of the Mini-Mult, the Hugo, the Midi-Mult, and the FAM

Requests for reprints should be sent to Norman G. Poythress, Jr., Center for Forensic Psychiatry, P.O. Box 2060, Ann Arbor, Michigan 48106.

was covered in Faschingbauer (1976) and will not be repeated here. The MMPI-168, developed by Overall and Gomez-Mont (1974), is the newest of the MMPI short forms. It consists of the first 168 items of the regular MMPI (the first 7 pages of the MMPI, Form R), which are scored by the regular scoring keys. These reduced MMPI raw scale scores are then converted to estimates of the full MMPI scale scores via either regression equations (Overall & Gomez-Mont, 1974, Table 1) or a conversion table (Overall, Higgins, & Schweinitz, 1976, Table 5). Early reports of the statistical concordance between the MMPI-168 and the full MMPI have been encouraging. Overall and Gomez-Mont reported correlations with the full MMPI scales ranging from .79 to .96, with a median correlation of .89 in a sample of 339 psychiatric patients. Newmark, Newmark, and Cook (1975) found that for male psychiatric patients, scale correlations between the MMPI-168 and the full MMPI ranged from .78 to .96, with a median of .88; for female psychiatric patients the range was from .77 to .93, with a median of .90. Further, in about 72% of cases the MMPI-168 predicted either the first or first and second peak scales found on the full MMPI. These initial findings for the MMPI-168 compare quite favorably with the correlations and concordance rates established for the other available short forms (Faschingbauer, 1976, Tables 1 and 2).

Statistical Concordance and Clinical Validity

For the practicing clinician who is looking for an MMPI short form appropriate for use in individual assessment, the relevance of statistical concordance data between various short forms

and the full MMPI must be questioned. Clearly, short-form to long-form scale correlations offer little of value concerning the utility of short-form profiles for individual profile interpretation and clinical decision making. The concordance for elevated peaks and code types is somewhat more relevant, and several investigators of the utility of MMPI short forms have offered negative evaluations of short forms based on concordance rates, which they judged to be unacceptably low. Hoffman and Butcher (1975), for example, found low hit rates for the Mini-Mult, the FAM, and the MMPI-168, and concluded that

there is insufficient evidence to advocate clinical use of the MMPI short forms. It seems that with such low classification accuracy in the short forms they would not simply plug into the existing interpretations and uses of the standard form without some modification and cautions. (p. 38)

Similarly, Griffin, Finch, and Edwards (1976) concluded their study of the Midi-Mult with the following warning: "The Midi-Mult cannot be used as an MMPI short form . . . because its scales are not sufficiently accurate in predicting the MMPI" (p. 56).

Other investigators have found these judgments and warnings to be premature or inappropriate. Poythress and Blaney (in press) noted:

This conclusion . . . suffers from three shortcomings. First of all, discordance of code type does not necessarily mean discordance of interpretation. One popular MMPI handbook (Gilberstadt & Duker, 1965) lists anxiety neurosis as an appropriate diagnosis for code types 1-3-9, 2-7, or 4, no two of which share even a single peak. Secondly, most investigators of code type concordance have concluded that the values they obtained were unacceptably low. However, if the code type concordance rate for test-retest studies of the full MMPI is utilized as the criterion, values such as those obtained by Hoffman and Butcher appear to be much more acceptable. For example, Faschingbauer (1974) found that for 61 subjects who took the full MMPI twice with only a one day interval, the same two-point code type was obtained in only 41% of cases. This suggests that the degree of slippage between long and short forms may not be markedly greater than the slippage between two long forms administered in close succession. Finally, the conclusion of Hoffman and Butcher and others is based on only an intermediate step in clinical interpretation. The final step, the clinician's profile interpretation, can be observed and measured directly; thus these investigators have made inferences where empirical investigation was called for. (p. 3)

The remainder of this comment is devoted to a brief review of empirical studies in which some

clinical judgment, decision, or interpretation was used to evaluate the utility of one or more of the MMPI short forms relative to the MMPI. It is asserted that this is perhaps the most relevant body of research literature for the practitioner to consider when contemplating the use of an MMPI short form.

Clinical Validity Studies: Diagnostics

Table 1 summarizes the findings of seven recent studies comparing one or more of the MMPI short forms to the full MMPI in diagnostic decision-making situations. In four of the studies shown, the consensual diagnoses of two or three PhD psychologists in their blind analysis of short- or long-form profiles are compared. These studies suggest that the FAM and MMPI-168 are about equal (near 80% agreement with the full MMPI) and are superior to two other short forms, the Midi-Mult and the Hugo. In three validity studies using various external diagnostic criteria, the MMPI-168 has been found slightly superior to the full MMPI in diagnostic validity. These studies suggest that for diagnostic purposes, the clinician should consider using the FAM or the MMPI-168 over the other available short forms, with a slight edge going to the MMPI-168.

Clinical Validity Studies: Profile Interpretation

To date there are but three studies that directly address the issue of accuracy of individual profile interpretation—Newmark, Conger, and Faschingbauer (1976); Newmark, Falk, and Finch (1976); and Poythress and Blaney (in press).

In the studies by Newmark and his associates, Newmark served as a blind interpreter of long-form and short-form profiles, generating 200- to 300-word interpretive statements about the patients whose profiles were provided. Subsequently, psychiatric residents in charge of patient care have rated his blind interpretive statements on a scale from 1 (totally inaccurate) to 5 (totally accurate), using their own knowledge of the patient as the criteria.

Newmark, Conger, and Faschingbauer (1976) found that the FAM compared favorably with the full MMPI, with 84% of the FAM interpretations and 92% of the full MMPI interpretations receiving a rating of either 4 (80% accurate) or 5 (totally accurate). In paired comparisons of long-form interpretations and short-form interpretations on the same patient, the full

Table 1
Diagnostic Validity of MMPI Short Forms

Study	Diagnostic outcome variable	Agreement between
Newmark, Cook, Clark, & Faschingbauer (1973)	Consensual diagnosis of 2 or 3 PhD psychologists—general categories of psychotic, neurotic, or personality disorder	MMPI and FAM: 76%
Newmark, Newmark, & Faschingbauer (1974)	same as above	MMPI and Midi-Mult: 43% MMPI and Hugo: 57% MMPI and FAM: 78%
Newmark, Owen, & Newmark (1975)	same as above	MMPI and Midi-Mult: 52% MMPI and Hugo: 64% MMPI and FAM: 78%
Newmark, Newmark, & Cook (1975)	same as above	MMPI and MMPI-168: 83%
Newmark & Finch (1976)	Same diagnostic decision as above for profiles, but hospital staff diagnosis utilized as criterion	MMPI & criterion: 84% MMPI-168 & criterion: 88%
Overall, Butcher, & Hunter (1975)	Using linear discriminant function to discriminate normals from psychiatric patients based on MMPI or MMPI-168 scores	MMPI-168 slightly superior to full MMPI discriminating normals from psychiatric patients
Overall, Higgins, & Schweinitz (1976)	Using multiple discriminant function for differential diagnosis across 10 major psychiatric diagnostic categories based on MMPI or MMPI-168 scores	MMPI-168 slightly superior to full MMPI in this differential diagnostic task

Note. MMPI = Minnesota Multiphasic Personality Inventory; FAM = Faschingbauer Abbreviated MMPI.

MMPI interpretations were more often rated above the FAM interpretations than was the reverse, but further analyses estimating validity coefficients (corrected for chance) for these two instruments yielded a coefficient of .94 for the full MMPI and .89 for the FAM. A second study, by Newmark, Falk, and Finch (1976), extended this same methodology to the FAM, the Hugo, and the MMPI-168. These investigators found that about 90% of the full MMPI interpretations received ratings of 4 or 5; the FAM, the Hugo, and the MMPI-168 received 82%, 45%, and 86%, respectively. The mean rating for the full MMPI interpretations was significantly higher than that for either the FAM or the Hugo but not for that of the MMPI-168. In paired comparisons, the MMPI interpretations were more often rated above the short-form interpretations, but this dominance was pronounced only for the Hugo. The MMPI-168 fared slightly better than the FAM and was judged to be, within the limits of the methodology of that study, interpretively equal to the full MMPI.

Poythress and Blaney (in press) noted the problems in the Newmark, Conger, and Fasch-

ingbauer (1976) and Newmark, Falk, and Finch (1976) methodology—the use of only one clinical interpreter and the gross measure of interpretive accuracy—and approached the assessment of short-form interpretive validity in a different manner. First, a large pool of clinical psychologists ($N = 29$) from various clinical and academic settings were recruited as blind interpreters of MMPI profiles. Second, profile interpretation was standardized and quantified by the use of a 30-item Q sort. For a given patient, four different raters were asked to rate profiles via the completion of a Q sort; two raters were mailed a profile based on full MMPI scoring, one rater received a Mini-Mult profile, and one rater received a FAM profile. The distribution of long-form/long-form Q -sort correlations was compared with each of the long-form/short-form correlation distributions to see if independent interpretations of the same long-form MMPI profiles were in fact more similar (i.e., correlated more highly) than were independent interpretations of a long form and a short form. The investigators found that the FAM was interpretively similar enough to the full MMPI to recommend it for clinical use; in 50% of com-

parisons, the FAM/MMPI Q-sort correlation was equal to or greater than the MMPI/MMPI Q-sort correlation for the same patient. However, the Mini-Mult did not generate interpretations sufficiently similar to those generated by the full MMPI to warrant a recommendation for its clinical use.

Discussion

Faschingbauer (1976) provided an extensive review of the literature on the ability of various short forms of the MMPI to predict various parameters of the full MMPI—such as elevated peaks or code types—and a discussion of differences in content among the various short forms. This article extends the review of Faschingbauer by providing a brief review of the empirical studies on the MMPI short forms in which an actual clinical decision or judgment was used as a dependent variable. It further includes a discussion of the recent MMPI-168, a short form that Faschingbauer's earlier article did not consider.

The needs and working constraints of the individual clinician will determine the selection of the MMPI short form. Unusual time constraints or illiterate subjects may dictate the use of one of the more abbreviated of the MMPI short forms, such as the Midi-Mult (86 items) or the Mini-Mult (71 items), which can easily be administered orally. The present review of validity studies, however, suggests that the FAM or the MMPI-168 should be selected over the other available short forms, since these two seem to provide the most accurate diagnostic and interpretive information. A slight edge in favor of the MMPI-168 is suggested, and it is the easiest of the short forms to administer. Form R of the regular MMPI can be used, with the patient being instructed to complete only the first 7 pages. Administration time is about 40 minutes, and conversion to full scale score estimates is easily accomplished via a conversion table.

References

- Dean, E. F. A lengthened Mini: The Midi-Mult. *Journal of Clinical Psychology*, 1972, 28, 68-71.
- Faschingbauer, T. R. A 166-item short form of the group MMPI: The FAM. *Journal of Consulting and Clinical Psychology*, 1974, 42, 645-655.
- Faschingbauer, T. R. Some clinical considerations in selecting a short form of the MMPI. *Professional Psychology*, 1976, 7, 167-176.
- Griffin, J. L., Finch, A. J., & Edwards, G. L. MMPI/Midi-Mult correspondence with parents of emotionally disturbed children. *Journal of Clinical Psychology*, 1976, 32, 54-56.
- Hoffman, N. G., & Butcher, J. N. Clinical limitations of three Minnesota Multiphasic Personality Inventory short forms. *Journal of Consulting and Clinical Psychology*, 1975, 43, 32-39.
- Hugo, J. Abbreviation of the Minnesota Multiphasic Personality Inventory through multiple regression. (Doctoral dissertation, University of Alabama, 1971). *Dissertation Abstracts International*, 1971, 32, 12B-13B. (University Microfilms No. 71-2011S)
- Kincannon, J. Prediction of the standard MMPI scale scores from 71 items: The Mini-Mult. *Journal of Consulting and Clinical Psychology*, 1968, 32, 319-325.
- Newmark, C. S., Conger, A. J., & Faschingbauer, T. R. The interpretive validity and effective test length functioning of an abbreviated MMPI relative to the standard MMPI. *Journal of Clinical Psychology*, 1976, 32, 27-32.
- Newmark, C. S., Cook, L., Clark, M., & Faschingbauer, T. R. Applications of Faschingbauer's Abbreviated MMPI to psychiatric inpatients. *Journal of Consulting and Clinical Psychology*, 1973, 41, 416-421.
- Newmark, C. S., Falk, R., & Finch, A. J. Interpretive accuracy of abbreviated MMPIs. *Journal of Personality Assessment*, 1976, 40, 266-268.
- Newmark, C. S., & Finch, A. J. Comparing the diagnostic validity of an abbreviated and standard MMPI. *Journal of Personality Assessment*, 1976, 40, 10-12.
- Newmark, C. S., Newmark, L., & Cook, L. The MMPI-168 with psychiatric patients. *Journal of Clinical Psychology*, 1975, 31, 61-64.
- Newmark, C. S., Newmark, L., & Faschingbauer, T. R. Utility of three abbreviated MMPIs with psychiatric outpatients. *Journal of Nervous and Mental Disease*, 1974, 159, 438-443.
- Newmark, C. S., Owen, M., & Newmark, L. Comparison of three abbreviated MMPIs for psychiatric patients and normals. *Journal of Personality Assessment*, 1975, 39, 261-270.
- Overall, J. E., Butcher, J. N., & Hunter, S. Validity of the MMPI-168 for psychiatric screening. *Educational and Psychological Measurement*, 1975, 35, 393-400.
- Overall, J. E., & Gomez-Mont, F. The MMPI-168 for psychiatric screening. *Educational and Psychological Measurement*, 1974, 34, 315-319.
- Overall, J. E., Higgins, W., & Schweinitz, A. D. Comparison of differential diagnostic discrimination for abbreviated and standard MMPI. *Journal of Clinical Psychology*, 1976, 32, 237-245.
- Poythress, N. G., & Blaney, P. H. The validity of MMPI interpretations based on the Mini-Mult and the FAM. *Journal of Personality Assessment*, in press.

Received August 31, 1977 ■

A Note on the MMPI as a Suicide Predictor

James R. Clopton
Texas Tech University

Comments are offered on a recent article by Leonard on the Minnesota Multiphasic Personality Inventory (MMPI) as a suicide predictor. Her appraisal of former studies and her consideration of the differences between individuals with different suicidal behaviors are critically evaluated. Future research should emphasize the development of useful MMPI indices of suicidal risk and should recognize the need to cross-validate these indices.

A recent article by Leonard (1977) regarding the Minnesota Multiphasic Personality Inventory (MMPI) as a suicide predictor deserves comment. MMPI data obtained from psychiatric inpatients who had committed suicide were compared with two matched control groups, one of highly suicidal patients and the other of non-suicidal patients. Discriminant analyses revealed that the committed-suicide MMPI profiles were distinguishable from nonsuicidal profiles. Leonard discussed the personality differences between suicidal and nonsuicidal patients and emphasized the lack of homogeneity among suicidal patients.

Despite mentioning some of the limitations of previous research on suicide, Leonard (1977) declared that this research provides "encouraging results." However, to date there has been no indication that standard MMPI scales, MMPI profile analysis, or specially developed MMPI suicide scales can *reliably* predict suicide at *useful* levels. Two review articles (Clopton, 1974; Lester, 1970) show that the results of previous research are inconsistent and reveal that previous studies often have had major methodological flaws, such as combining individuals with different types of suicidal behavior into the same group (e.g., Broida, 1954). Leonard also appears to have overlooked research studies that have provided negative results. For example, Ravensborg and Foss (1969), using a multivariate analysis, found no differences in the MMPI profiles of patients who had committed suicide in a state hospital, patients who died of natural causes in the same hospital, and nonsuicidal patients.

Leonard's (1977) study and two earlier studies (Clopton & Jones, 1975; Devries & Farberow, 1967) have found some evidence that multivariate analyses can differentiate suicidal from nonsuicidal patients. Correct classification was obtained for 41% of the patients in the Devries and Farberow (1967) study and for 66% of the patients in the Clopton and Jones (1975) study. Leonard reported that for the female patients in her study, a discriminant function using seven MMPI scales correctly classified all 16 women who committed suicide and 14 of 16 female patients in a comparison group. Close examination reveals that these results may be less impressive than they appear. As Leonard explained, the clinician wants a single index or other straightforward means of determining suicidal risk from MMPI data, and the variables in a discriminant analysis form a complex cluster. None of the three studies using multivariate procedures have examined the usefulness of these procedures to clinicians faced with the task of predicting suicidal behavior. Another consideration not mentioned by Leonard is that before multivariate procedures can be considered reliable in classifying patients as suicidal or nonsuicidal, they need to be cross-validated. That is, the combination of differentiating variables needs to be derived in one comparison of suicidal and nonsuicidal individuals and then extended to new samples. The percentage of correct classifications could be considerably lower in the validation samples. None of the three studies using multivariate procedures have cross-validated their procedures.

Leonard (1977) stressed the diversity existing among different suicidal groups, and at one point she stated that increasing the sample size in suicide studies by including suicide threateners or attempters with those who commit suicide is appealing but not justified. She stated that the populations (actual suicides, suicide attempters,

The author thanks Roger Greene and Donald H. Baucom for their helpful suggestions.

Requests for reprints should be sent to James R. Clopton, Department of Psychology, Texas Tech University, Lubbock, Texas 79409.

and suicide threateners) are obviously different. One is puzzled then to find that the high-suicide comparison group in her study contained patients with histories of suicide attempts and/or prolonged threats and preoccupation. Combining patients who threaten suicide with those who make suicide attempts is a questionable procedure. Previous MMPI research reveals that patients threatening suicide have the most deviant MMPI profiles of any suicidal group and that there are differences in the MMPI profiles of patients who threaten suicide and those who attempt suicide (Farberow, 1956; Farberow & Devries, 1967; Rosen, Hales, & Simon, 1954; Simon & Gilberstadt, 1958).

Leonard (1977) has perpetuated the notion that the study of suicide attempters is not a valid way to gain information about individuals who commit suicide. She stated that the two populations are different, and in support she noted that approximately three times more females than males attempt suicide, but approximately three times more males commit suicide. Despite these sex differences, the prevailing view is that suicide attempters and those who commit suicide can be regarded as members of the same population (Stengel, 1972). Chance often plays a critical role in influencing the outcome of a suicide attempt, regardless of the lethality of the attempt. Also, those who commit suicide frequently have prior records of unsuccessful attempts. In Leonard's committed suicide group, 80% of the males and 75% of the females had a history of one or more suicide attempts prior to committing suicide. Finally, observed sex and age differences between those who commit suicide and those who attempt suicide may be largely due to incidental cultural factors. The higher rate of committed suicide among males is probably related to their use of more violent methods, such as gunshot, which provide less opportunity for rescue than self-poisoning, the preferred method of females. Because men have more access to instruments of violence and more training in their use, it is reasonable to expect more of their suicide attempts to be successful.

Leonard's (1977) study is a valuable contribution to the research literature, but neither her results nor the findings of previous studies have indicated that the MMPI can reliably predict suicide at useful levels. Multivariate procedures, such as those used by Leonard, have not been cross-validated, nor have they been demonstrated to be of practical value to clinicians attempting to predict suicidal behavior.

Leonard (1977) combined patients who threatened suicide and patients who attempted suicide

into one group. However, patients who threaten suicide are the most distinct suicide group. It is possible that the patients who attempted suicide were more similar to the patients who committed suicide than to the patients who threatened suicide. Despite Leonard's comments to the contrary, the current view is that patients who attempt suicide are probably all members of the same population regardless of the success of their attempts.

References

- Broida, D. G. An investigation of certain psychodiagnostic indications of suicidal tendencies and depression in mental hospital patients. *Psychiatric Quarterly*, 1954, 28, 453-464.
- Clopton, J. R. Suicidal risk assessment via the Minnesota Multiphasic Personality Inventory (MMPI). In C. Neuringer (Ed.), *Psychological assessment of suicidal risk*. Springfield, Ill.: Charles C Thomas, 1974.
- Clopton, J. R., & Jones, W. C. Use of the MMPI in the prediction of suicide. *Journal of Clinical Psychology*, 1975, 31, 52-54.
- Devries, A. G., & Farberow, N. J. A multivariate profile analysis of MMPIs of suicidal and non-suicidal neuropsychiatric hospital patients. *Journal of Projective Techniques and Personality Assessment*, 1967, 31, 81-84.
- Farberow, N. L. Personality patterns of suicidal mental hospital patients. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: University of Minnesota Press, 1956.
- Farberow, N. L., & Devries, A. G. An item differentiation analysis of suicidal neuropsychiatric hospital patients. *Psychological Reports*, 1967, 20, 607-617.
- Leonard, C. V. The MMPI as a suicide predictor. *Journal of Consulting and Clinical Psychology*, 1977, 45, 367-377.
- Lester, D. Attempts to predict suicidal risk using psychological tests. *Psychological Bulletin*, 1970, 74, 1-17.
- Ravensborg, M. R., & Foss, A. Suicide and natural death in a state hospital population: A comparison of admission complaints, MMPI profiles, and social competence factors. *Journal of Consulting and Clinical Psychology*, 1969, 33, 466-471.
- Rosen, A., Hales, W. M., & Simon, W. Classification of "suicidal" patients. *Journal of Consulting Psychology*, 1954, 18, 359-362.
- Simon, W., & Gilberstadt, H. Analyses of the personality structure of 26 actual suicides. *Journal of Nervous and Mental Disease*, 1958, 127, 555-557.
- Stengel, E. A survey of follow-up examinations of attempted suicides. In J. Waldenstrom, T. Larsson, & N. Ljungstedt (Eds.), *Suicide and attempted suicide*. Stockholm: Nordiska Bokhandeln's Forlag, 1972.

Received July 26, 1977 ■

Response to "A Note on the MMPI as a Suicide Predictor"

Calista V. Leonard
University of California, Los Angeles

Response is made to comments by Clopton on an article by Leonard on the Minnesota Multiphasic Personality Inventory (MMPI) as a suicide predictor. Clopton's viewpoint that patients who attempt suicide are probably all members of the same population regardless of the success of their attempts is discussed. His concern that future research should emphasize the development of useful MMPI indices of suicidal risk and should recognize the need to cross-validate these indices is endorsed.

Clopton's (1978) main concern in his comments on Leonard (1977) seems to center about his belief that "patients who attempt suicide are probably all members of the same population regardless of the success of their attempts" (p. 336). This is an individual viewpoint rather than a "current view." It does not take a statistician to deduce that if attempted suicides are all from the same population, then one would expect male suicides, for example, to be equaled by female suicides instead of being approximately three times more numerous. Unquestionably, there is overlapping between the groups, but many suicides have never attempted or overtly threatened suicide (Robins, Gassner, Kayes, Wilkinson, & Murphy, 1959), and many attempted suicides who have seemed lethally suicidal do not go on to commit suicide (Pitts & Winokur, 1964).

Another concern expressed by Clopton is that negative results were not reported, and I agree in this respect. However, negative results must be studied for comparability, and the Ravensborg and Foss (1969) study that he cites (Clopton, 1978) contained control groups that were not matched with the committed suicide group on important variables. The control groups were significantly older, less educated, and had lower intelligence scores ($p < .001$). One must wonder why the multivariate analysis found no significant differences in Minnesota Multiphasic Personality Inventory (MMPI) profiles in spite of these important group differences.

A further criticism with which I agree wholeheartedly is that cross-validation and helpful

hints for clinicians are sorely needed in studying the MMPI as a predictor for suicide. Even though Clopton (1978) did not report a cross-validation for his study (Clopton & Jones, 1975), his discriminant analysis of the data lent some support to the use of MMPI profiles to identify suicidal psychiatric patients. His caution in this excellent study concerning the need for identifying the obtained differences in profiles is well taken.

I would like to conclude this response with a note on armchair theorizing, which I enjoy and define as "making hypotheses without providing evidence for them" and which seems to me to be the forerunner of attempts at finding testable answers. Clopton (1978) theorizes that men commit suicide more frequently because they have more access to instruments of violence and more training in their use, whereas women are less frequently suicides because they use self-poisoning and are more readily rescued. The immediate response, of course, is "but any ambulatory and determined would-be suicide could select jumping as a method, and most women drive and could use a car lethally even if they didn't have training or access to guns." Which brings the issue directly back to the cultural and personality variables that lead one person to choose a gun, another to make a half-hearted pill-taking attempt, and another to choose a nonsuicidal solution to environmental stresses. This is what research on the prediction of suicide is all about. What environmental and personality differences make one person more self-lethal than another?

References

- Clopton, J. R. A note on the MMPI as a suicide predictor. *Journal of Consulting and Clinical Psychology*, 1978, 46, 335-336.

Requests for reprints should be sent to Calista V. Leonard, Department of Psychiatry, University of California, 760 Westwood Plaza, Los Angeles, California 90024.

- Clopton, J. R., & Jones, W. C. Use of the MMPI in the prediction of suicide. *Journal of Clinical Psychology*, 1975, 31, 52-54.
- Leonard, C. V. The MMPI as a suicide predictor. *Journal of Consulting and Clinical Psychology*, 1977, 45, 367-377.
- Pitts, F. N., & Winokur, G. Affective disorder: III. Diagnostic correlates and incidence of suicide. *Journal of Nervous and Mental Disease*, 1964, 139, 176-181.
- Ravensborg, M. R., & Foss, A. Suicide and natural death in a state hospital population: A comparison of admission complaints, MMPI profiles, and social competence factors. *Journal of Consulting and Clinical Psychology*, 1969, 33, 466-471.
- Robins, E., Gassner, J., Kayes, J., Wilkinson, R. H. Jr., & Murphy, G. E. The communication of suicidal intent: A study of 134 cases of successful (completed) suicide. *American Journal of Psychiatry*, 1959, 115, 724-733.

Received October 19, 1977 ■

Brief Reports

Characteristics of the Emotional Responsiveness of Sensitizers and Repressors to Social Stimuli

Glenn A. Miller
Arizona State University

William Nuessle
University of Kentucky

TAT-like (Thematic Apperception Test) stimuli of positive, negative, and neutral affect were presented to 30 sensitizers, 30 intermediates, and 30 repressors to test Lefcourt's formulation that sensitizers positively value emotional expression and repressors devalue such expression. Contrary to the expectation based on an extension of Lefcourt's model, sensitizers did not significantly differ from repressors on the number of positive, negative, or total affect words in their stories. His model appears consistent only with the responses of subjects to negative affective stimuli and not with their responses to positive affective stimuli.

The traditional interpretation of repression and sensitization as referring to defensive dispositions to anxiety cues was challenged by Lefcourt (1966). He suggested that sensitizers and repressors differ in their valuation of emotional expression, with sensitizers positively and repressors negatively valuing such expression. For example, both repressors and sensitizers describe themselves in positive ways. The self-descriptions of sensitizers, which have been typically interpreted as negative, are actually positive from their viewpoint because they perceive their "admissions [of emotionality] as revealing honesty with one's self, and a lack of fear of self-disclosure" (Lefcourt, 1966, p. 445).

This reformulation minimizes the importance of anxiety cues and increases the relevance of the emotional expressiveness of repressors and sensitizers to positive and neutral stimuli. The present study attempted to determine whether Lefcourt's theory would predict the emotional expressiveness of repressors and sensitizers to positive and neutral as well as to negative stimuli.

The subjects, 90 male introductory psychology students, were divided on the basis of Byrne's Repression-Sensitization scale (Byrne, 1961) into three equal groups of sensitizers, intermediates, and repressors and were presented with eight TAT-like (Thematic Apperception

Test) slides to which they wrote stories. One third of each group received stimuli of positive, neutral, or negative affective tone. The dependent measures were the number of positive, of negative, and of total affective words contained in the stories.

The clearest test for the extension of Lefcourt's interpretation to both positive and negative emotional stimuli is the comparison of the three subject groups on the total number of emotional words to all three types of stimuli combined, with differences in length of story covaried out. Contrary to Lefcourt's theory, the sensitizers did not exceed repressors and intermediates on negative, positive, or total affective words. An interaction between the affective tone of the stimuli and the three subject groups occurred in the analysis of both the total emotional words and total negative words, $F(4, 71) = 2.82, p < .05$; $F(4, 71) = 2.50, p < .05$. The interaction for total emotional words is a consequence of the fact that sensitizers exceeded repressors on the negative slides but repressors gave more emotional words to the positive and neutral slides.

In the analysis of total positive words, there were no significant differences between groups and there was no interaction between subject groups and types of stimulus slides. In fact, sensitizers gave fewer positive words than repressors. This difference in means is opposite in direction, although not significantly, to that which would be predicted from Lefcourt's interpretation.

Requests for reprints should be sent to Glenn A. Miller, Department of Psychology, Arizona State University, Tempe, Arizona 85281.

Thus the greater emotional expressiveness of sensitizers, as compared with repressors, appears to be limited to responses to negative stimuli. This would seem to mean that Lefcourt's interpretation of the repression-sensitization dimension as a reflection of differential valuation of emotionality requires additional constructs to account for the interaction between the affective tone of the stimuli and the status of subjects as repressors and sensitizers.

References

- Byrne, D. The repression-sensitization scale: Rationale, reliability and validity. *Journal of Personality*, 1961, 29, 334-349.
- Lefcourt, H. M. Repression-sensitization: A measure of the evaluation of emotional expression. *Journal of Consulting Psychology*, 1966, 30, 444-449.

Received January 3, 1976 ■

Factors Influencing Lower-Class Black Patients Remaining in Treatment

Anthony Vail

San Diego County Mental Health, San Diego, California

Early termination from individual therapy in a community mental health clinic was studied with lower-class black patients assigned to therapists who were black or white. The only significant correlate was the interaction between sex of therapist and sex of patient. Patients remained longer with therapists of the opposite sex. No significant correlations were found between remaining in treatment and black patients' attitudes toward whites, patients' perceptions of therapists' understanding and acceptingness, or patient-therapist discrepancies in their perception of therapy.

This study explored early termination from psychotherapy in an inner-city community mental health clinic in Philadelphia. The 32 black male and 55 black female lower-class patients for this study were obtained after they had had clinical evaluations. They were randomly assigned to 1 of 10 staff therapists who were black or white, male or female. After the first therapy session, the therapists filled out the discrepancy scale described below. The patients were orally presented with this and other scales by an experimental interviewer.

Attitudes toward whites were assessed by an attitude behavior scale of blacks toward whites. Two levels of this scale were used that examined the patients' personal behavior and their attitudes toward societal stereotypes.

Patient-therapist discrepancies in perception of therapy were measured by a scale constructed for this study. It included an inquiry as to what the patients considered their most important problem; questions as to whether the therapy matched the patients' expectations; a sampling of the patients' attitudes toward seeking professional, psychological help; estimates of the therapists' understanding and acceptingness as perceived by the patients; and questions about the patients' preference for certain types and degrees of therapist activity.

Finally, patients' perceptions of their therapists' understanding and acceptingness were measured in detail.

The 43 patients who came for three sessions or less were considered the dropouts; the 44 who came for four or more sessions were considered

remainers. To analyze for effects of patient-therapist similarity in race and sex on continuation in treatment, a three-dimensional contingency table was used.

The only result significant at the .05 level was the Sex of the Therapist \times the Sex of the Patient interaction. The race of the therapist did not matter statistically. Most of the interaction originated with male therapists being less effective with black male patients than with black female patients. Female therapists were also less effective with patients of their own sex, although this was only about half as pronounced as the finding for male therapists. The interaction effect was opposite to the direction predicted. At least 29% of the variance among patients' remaining in treatment was accounted for by this interaction effect.

No significant correlations were found between remaining in treatment and the black patients' attitudes toward whites on either of two levels; patients' perception of their therapists' understanding and acceptingness; or the similarity of their view of therapy to that of their therapists.¹ In all of these analyses, it did not matter statistically whether the therapists were black or white.

It seems worth looking further into the dynamics of the cross-sex phenomenon between therapists and lower-class patients to see if it is based on the attractiveness of a heterosexual encounter, inhibitions associated with being painfully revealing of oneself before a therapist of the same sex, or some other reason.

Received February 12, 1976 ■

Requests for reprints and for an extended report of this study, and for a copy of the scales used, should be sent to Anthony Vail, 12732 Gibraltar Drive, San Diego, California 92128.

¹ A discussion of some small but significant correlations between Patient-Therapist Discrepancy subscales and dropout can be found in the extended report.

Effects of Explanation and Information Feedback on the Illusory Correlation Phenomenon

Ronald W. Waller and Stuart M. Keeley
Bowling Green State University

An attempt was made to attenuate the illusory correlation phenomenon existent in one set of diagnostic stimulus materials, Draw A Person, through training on another set, Rorschach. Three training methods were tried: information feedback plus explanation, information feedback only, and explanation. None of the training conditions attenuated the illusory correlation effect when no relationship existed in the stimulus set. Significant attenuation occurred only when a negative correlation existed.

Even though it seems clear that clinical psychologists have considerable room for improvement in making judgments, only recently have researchers begun to isolate factors contributing to nonoptimality of judgment and to the frequently misplaced faith in such judgments. One such factor is the presence of "illusory correlations"; that is, subjects with strong *a priori* expectations often respond to expected relationships rather than to the relations actually in the stimulus materials (e.g., Chapman & Chapman, 1969). This tendency to maintain a belief in illusory correlates when the evidence does not support such a belief presents a major obstacle to making valid clinical judgments. This study asks the question, "Can this obstacle be overcome by sensitizing judges to the phenomenon?" Such training of judges would have maximum utility if it were to generalize to a variety of judgment tasks. The present study examines whether special training on illusory correlation using Rorschach test stimulus materials will reduce the illusory correlation effect for a second set of materials, the Draw-A-Person (DAP) task.

Subjects were 120 volunteer undergraduate introductory psychology students. Thirty subjects were randomly assigned to each of four different training conditions: (a) explanation, (b) information feedback, (c) explanation and information feedback, and (d) control. Subjects assigned to each of the four training conditions were divided into three subgroups ($n = 10$), each receiving a different set of generalization stimulus materials. Thus, there were 12 groups in all. Each group participated in three phases: (a)

pretraining judgments, (b) training, and (c) posttraining generalization judgments.

In the pretraining phase, subjects made judgments about how they believed specified Rorschach and DAP cues were related to various symptoms. In the training phase, subjects in each training group were exposed to the same three sets of 24 Rorschach stimuli each. Each set included one cue-symptom pair that had been found to be an illusory correlate in previous research or else had been rated as being related by more than 50% of a group of 25 introductory psychology pilot subjects. Set 1 contained no relationship among cues and symptoms. Set 2 contained a strong negative relationship between the illusory correlate pair; and Set 3 contained a strong positive relationship.

Subjects in the information feedback and explanation training condition (FE) were given an explanation of the nature of the illusory correlation phenomena and were informed of research findings related to this phenomenon prior to viewing the three training sets. They were warned that illusory relationships might bias their own judgments of the stimulus materials and were cautioned to try to avoid this kind of bias. After viewing each stimulus set and judging the cue-symptom relationships, subjects were given the actual cue-symptom relationships contained in that set to compare with their predictions. Subjects in the explanation (E) condition received a training procedure identical to the FE group, except they received no information concerning the actual relationships contained in the stimulus materials. Subjects in the information feedback (F) condition were only given feedback information about the true cue-symptom relationships. Control subjects (C) received no information concerning the actual

Requests for reprints should be sent to Stuart M. Keeley, Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403.

cue-symptom relationships and no explanation of the illusory correlation effect.

In the generalization judgment phase, each subgroup was exposed to one of three sets of DAP stimulus materials, each set varying in actual cue-symptom relationships. Stimuli were similar to those used by Chapman and Chapman (1967). Each set contained the illusory correlate of broad shoulders and muscular drawing characteristics and the symptom "he is worried about how manly he is." The three cue-symptom relationships were (a) no relationship among DAP drawing characteristics and the symptom statements (DS 1), (b) a strong negative relationship between the illusory correlate pair (DS 2), and (c) a strong positive relationship between the illusory correlate pair (DS 3).

Subjects made two kinds of responses during each phase. First, they used a 7-point scale indicating the likelihood that the patient had each symptom given a specified cue. Second, they rated the confidence they had in the judgments they made about the relationships in each stimulus set by marking a 15-cm line.

As expected, a strong a priori bias existed in the pretraining estimates of the relationship between the illusory cue-symptom pairs. The majority of subjects in all groups reported positive relationships for both Rorschach and DAP illusory pairs.

Likelihood ratings of the illusory DAP cue-symptom combination were analyzed by a three-way fixed effects analysis of variance (4 Training Conditions \times 3 Stimulus Sets \times 2 Judgment Times), with the subject factor nested under both training conditions and stimulus sets. Main effects were found for both stimulus set ($p < .01$) and time of judgment ($p < .01$); and a significant interaction was found between stimulus set and time ($p < .01$). The source of the interaction is clear. Subjects behaved in a virtually identical fashion on pretraining and posttraining judgments to DS 1 and DS 3 sets, reporting a positive relationship. However, subjects in all training conditions responded very differently to the DS 2 stimulus set following training than prior to training. Their judgments shifted from positive to either zero or negative. This differential shift is supported by the findings of a simple main effect of stimulus set for the second set of judgments ($p < .01$) and none for the first set.

A three-way analysis of variance on confidence judgments showed only one significant effect—time of judgment ($p < .01$). Subjects became more confident in their judgments from pretraining to posttraining, regardless of which training condition they were in and regardless of which stimulus set they were exposed to.

Sensitizing subjects to the illusory correlation phenomenon or Rorschach materials through feedback about actual relationships, explanation of the phenomenon, or both did not attenuate the tendency to see positive cue-symptom relationships when no relationship actually existed. However, negative relationships between illusory correlates markedly attenuated the effect. Such findings suggest that subjects do attend to the evidence. They also suggest that the number of confirmations of the original biased expectancy relative to the number of disconfirmations may be a crucial determinant of whether the original belief is maintained.

It is important to note that in the typical illusory correlation demonstration, subjects are provided with only affirmations of the symptoms; negations must be implied. Thus, they typically see some instances of cue present, symptom present for the illusory pair but never see the cue with an explicit statement of the absence of the symptom. Only a few affirmatives may be necessary to maintain the illusory belief.

The present study provides further suggestive evidence that human judges have a difficult time using evidence appropriately in the face of uncertainty. Short-term explanation and/or feedback procedures do not seem to be a sufficient remedy for this particular kind of error in human judgment. Whether long-term training procedures would attenuate the effect is presently an empirical question.

References

- Chapman, L. J., & Chapman, J. P. The genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, 1967, 72, 193-204.
- Chapman, L. J., & Chapman, J. P. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 1969, 74, 271-280.

Received September 7, 1976 ■

The 4-3 MMPI Profile Type: A Failure to Replicate

Jeffrey A. Buck and John R. Graham
Kent State University

Using a sample of 65 prison inmates, the incidence of violent crimes for persons with the 4-3 Minnesota Multiphasic Personality Inventory profile type and with other two-point code types was compared. The results failed to replicate the findings of some earlier investigators, who reported that violent behaviors are more common for persons with the 4-3 profile type. The failure to identify a significant relationship between the 4-3 profile type and violent behavior suggests that caution should be exercised in generalizing to populations that differ from those in which the relationship between violence and the 4-3 profile type is established.

Previous investigations of the Minnesota Multiphasic Personality Inventory 4-3 profile type have yielded inconsistent results. Some studies (e.g., Davis & Sines, 1971; Persons & Marks, 1971) have indicated that subjects producing the 4-3 profile are prone to act in more aggressive and violent ways than subjects not producing the 4-3 profile. Other studies (e.g., Gynther, Altman, & Warbin, 1973) have not found violence or hostility to be characteristic of persons with the 4-3 profile. Two explanations offered for this disparity have been differences in the ages of the subjects studied and the ways in which the 4-3 profile type have been defined. The present study attempted to gain information about both of these possible sources of variance.

Subjects were male inmates of a medium-security penitentiary for adult male felons located at Marion, Ohio. The incidence of violent crime was examined in a sample of 65 prisoners producing a 4-3 profile type and 64 not of this type. Violent crime was defined as murder, rape, aggravated assault, or robbery. The 4-3 and non-4-3 groups were separated into "old" subjects (30 or older) and "young" subjects (under 30). Within each of the resultant subgroups, the number of subjects committing violent and nonviolent crimes was determined and a chi-square test was made to determine if the subgroups differed on this variable. In addition to this over-

all comparison, different combinations of the subgroups were examined for possible differences on the violence measure. Analyses of variance also were performed to determine if the absolute and relative elevations of Scales 3 and 4 influenced the relationship between the 4-3 profile type and violent or aggressive behavior. All of the above tests were performed separately for black and white subjects. None of the relevant comparisons was significant.

Comparisons of the methodology of the present study with those of previous ones indicate that the way in which the 4-3 profile has been defined and the criterion measures of aggressive and violent behavior cannot completely account for the differences in findings between this investigation and others. A possible explanation for these differences is that the subjects in this study came from a facility in which the most aggressive and violent prisoners already had been selected out and sent elsewhere. The failure to identify a relationship between the 4-3 profile type and aggressive and violent behavior in this study suggests that caution should be exercised in generalizing to populations that differ from those in which the relationship between violence and the 4-3 profile type is established.

References

- This article is based on a thesis by the first author and directed by the second author, submitted in partial fulfillment of the requirements for the MA degree at Kent State University. The cooperation of the Ohio Department of Rehabilitation and Correction in the preparation of this study is gratefully acknowledged.
- Requests for reprints and for an extended report of this study should be sent to Jeffrey A. Buck, Department of Psychology, Kent State University, Kent, Ohio 44242.
- Davis, K. R., & Sines, J. O. An antisocial behavior pattern associated with a specific MMPI profile. *Journal of Consulting and Clinical Psychology*, 1971, 36, 229-234.
- Gynther, M. D., Altman, A., & Warbin, R. W. A new actuarial-empirical automated MMPI interpretive program: The 4-3/3-4 code type. *Journal of Clinical Psychology*, 1973, 29, 229-231.
- Persons, R. W., & Marks, P. A. The violent 4-3 MMPI personality type. *Journal of Consulting and Clinical Psychology*, 1971, 36, 189-196.

Received October 6, 1976 ■

Congruence of Parental Perception, Marital Satisfaction, and Child Adjustment

Lucy Rau Ferguson and Deborah R. Allen
Michigan State University

Fathers and mothers of 95 children 5-7 years old completed the Locke-Wallace Scale, the Interpersonal Checklist, and the Children's Behavior Checklist to assess marital satisfaction, congruence of self- and mate-perceptions, and agreement in parents' perceptions of their child and child adjustment, respectively. All variables were significantly, positively intercorrelated. Strongest association was between congruence in parents' perceptions of the child and child adjustment. Similarity in partners' self-concepts and psychological empathy were significantly associated with marital satisfaction and child adjustment. A general dimension of family harmony (vs. conflict) is seen as contributing to children's social adjustment.

Children who exhibit deviant social behavior generally come from families characterized by marital dissatisfaction and discrepant parental attitudes. In a recent study based on small samples (Ferguson, Partyka, & Lester, 1974), parents of well-adjusted children showed closer agreement in their perceptions of their child than parents of clinic-referred children. It was predicted that in the general population, parents who described their children as more favorably adjusted would show closer agreement in their perceptions of their child, would report greater marital satisfaction, and would show greater agreement between descriptions of themselves and descriptions of them by their spouses.

Letters requesting cooperation in a research study were distributed to all parents of kindergarten, first-, and second-grade children in the Holt, Michigan, Public School System (approximately 1,000 children). Two hundred four families returned postcards indicating their willingness to participate in the study. Parents of 95 children (51 males and 44 females between 5 and 7 years old) completed the following questionnaires: Marital adjustment and satisfaction were assessed for each parent by his/her responses to the Locke-Wallace Scale (Locke & Wallace, 1959). The Interpersonal Checklist (LaForge & Suczek, 1955) was used to assess parents' perceptions of self and spouse. The Children's Behavior Checklist (CBCL) was the in-

strument used to assess parents' perceptions of their child. It consists of 154 interpersonal and symptomatic items referring to the behavior of children; the parent checks whether each item is applicable to the child and/or characteristic of him/her. The CBCL was also used as a measure of child adjustment. Sixty-six items have been found to discriminate significantly between clinic-referred and nonclinic (adjusted) children, with 32 being more characteristic of clinic-referred children and 34 being more characteristic of nonclinic children (Ferguson et al., 1974). Adjustment is calculated by subtracting each parent's score on the 32 clinic items from his/her score on the 34 nonclinic items; the mean of the mother's and father's scores yields a combined adjustment score for the child.

Interitem phi coefficients expressed the agreement between the husband's perception of himself and his wife's perceptions of him and between the wife's perceptions of herself and her husband's perceptions of her. Agreement in parents' perceptions of their child was determined by calculating the correlations between the mother's ratings of her child on the CBCL and those of the father.

Distributions of Locke-Wallace scores and of CBCL child adjustment scores were all negatively skewed; thus this is a sample of families who preponderantly reported themselves as happy and well-adjusted. Associations between the various measures of congruence of parental perception (of spouse and child), marital satisfaction, and child adjustment were examined by means of product-moment correlation coefficients. For the sexes combined all correlations were significant at the .05 level in the predicted direc-

Requests for reprints and for an extended report of this study should be sent to Lucy Rau Ferguson, Department of Psychology, Michigan State University, East Lansing, Michigan 48824.

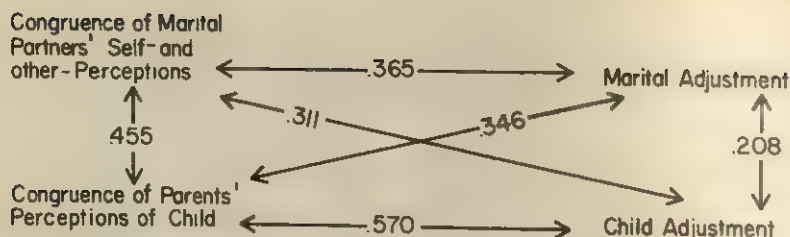


Figure 1. Intercorrelations of variables based on combined sample.

tion; 9 out of 13 were significant at the .001 level.

Figure 1 shows the interrelations among the variables.

Parents' agreement in viewing their child was closely associated with congruence in their perceptions of each other, and each of these variables in turn was significantly related to marital satisfaction. Thus we seem to be tapping a more general dimension of family harmony that underlies the child's social adjustment. When parents see their child as possessing the characteristics of well-adjusted children, they also tend to agree closely in their perceptions of all aspects of their child's behavior, to express satisfaction with their marriage, and to see their spouses the way their spouses see themselves. Further research is

underway to confirm these relationships using an independent measure of child adjustment.

References

- Ferguson, L. R., Partyka, L. B., & Lester, B. M. Patterns of parent perception differentiating clinic from nonclinic children. *Journal of Abnormal Child Psychology*, 1974, 2, 169-181.
- LaForge, R., & Suczek, R. The interpersonal dimension of personality: III. An interpersonal checklist. *Journal of Personality*, 1955, 24, 94-112.
- Locke, H. J., & Wallace, K. M. Short marital adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living*, 1959, 21, 251-255.

Received November 4, 1976 ■

A Factor Analysis of Assertive Behaviors

Joseph S. Pachman, David W. Foy, and Frank Massey
Veterans Administration Center, Jackson, Mississippi

Richard M. Eisler
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Five objective measures of interpersonal behavior presumed to be tapping "assertiveness" were factor analyzed. In addition, each behavioral measure was correlated with a subjective global rating of assertiveness. Four of the behavioral measures loaded highly on a general factor of assertiveness. The fifth behavioral measure loaded highly on a separate factor, Response Latency. All behavioral measures with the exception of response latency evidenced significant correlations with the subjective rating of global assertiveness.

Systematic progress in the area of assertiveness training has been hampered by a general lack of consensus as to what specific responses are involved in behaving "assertively." Some workers have chosen to define assertive behaviors intuitively and then modify these components (Wolpe & Lazarus, 1966). More recently, Eisler, Miller, and Hersen (1973) have identified specific verbal and nonverbal behaviors that distinguished psychiatric patients judged to be relatively assertive from those who were unassertive.

The present study was designed to statistically investigate the relative contributions of these presumed behavioral components of assertiveness with a factor-analytic procedure (principal axis method, varimax rotation). A second purpose was to compare subjective global ratings of assertiveness with specific behavioral measures.

The subjects were 55 males hospitalized for treatment of alcoholism. We constructed a series of five vocally related interpersonal encounters requiring assertive responses. A male respondent was employed to prompt subjects' responses to these situations. Responses were videotaped and rated on the following components of assertion: duration of eye contact, duration of reply, latency of reply, statements indicating "compliance" with the respondent's unreasonable requests and assertive requests for the respondent to change his behavior, and a global measure of assertiveness.

The results indicated that duration of reply (.69), duration of eye contact (.70), noncompliant speech content (.74), speech content requesting new behavior (.78), and overall asser-

Table 1

Correlations Between Global Assertiveness and Behavioral Components of Assertiveness

Behavior component	r
Response duration	.27*
Response latency	-.17
Request for change	.57**
Compliance	-.50**
Eye contact	.48**

* $p < .05$.

** $p < .001$.

tion (.75) were all loaded highly on Factor 1 (Assertiveness). However, latency of reply loaded nearly entirely (.93) on a separate factor (Factor 2, Response Latency). In addition, correlation coefficients obtained between the subjective global measure of assertiveness and the other behavioral measures, with the exception of latency of reply, were statistically significant.

The results corroborate the importance of the previously identified component behaviors (Eisler et al., 1973), with the exception of response latency, in reflecting assertiveness (see Table 1). The utility of a subjective global rating of assertiveness by trained observers was also demonstrated in view of the significant correlation coefficients obtained between this measure and the objective behavioral measures of assertion.

References

- Eisler, R. M., Miller, P. M., & Hersen, M. Components of assertive behavior. *Journal of Clinical Psychology*, 1973, 29, 295-299.
Wolpe, J., & Lazarus, A. A. *Behavior therapy techniques: A guide to the treatment of neuroses*. New York: Pergamon Press, 1966.

Received January 6, 1977 ■

Requests for reprints should be sent to Richard M. Eisler, Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

Life Change, the Sensation Seeking Motive, and Psychological Distress

Ronald E. Smith, James H. Johnson, and Irwin G. Sarason
University of Washington

The relationship between life change and psychological distress as a function of the sensation seeking motive was investigated. Scores on Lanyon's Discomfort scale were unrelated to positive and total life change scores but were significantly related to amount of negative life change occurring over the previous year. However, as predicted, this relationship was restricted to subjects low in sensation seeking; high sensation seekers are apparently more tolerant of negative life change.

The role of life change in physical and psychological well-being has received considerable empirical attention in recent years, and degree of life change has been shown to be significantly related to both physical illness and measures of psychopathology. Most of the life stress research has been based on the possibly erroneous assumption that change per se is stressful, regardless of whether it is perceived by the individual to be positive or negative. In addition, the role of personality variables as mediators of responses to life change has been virtually ignored.

The present study examined the relationships between both positive and negative life change and a measure of psychological distress as a function of subjects' scores on the Sensation-Seeking Scale (SSS; Zuckerman, Kolin, Price, & Zoob, 1964). It was hypothesized that low sensation seekers, who presumably have a low optimal level of stimulation, would be more negatively affected by stressful life changes than would high sensation seekers.

Forty-two male and 33 female college undergraduates were administered the 22-item SSS and the Life Experiences Survey (Sarason & Johnson, 1976). On the latter measure, subjects indicate which of a series of life changes they have experienced during the past 12 months, and they rate the positiveness or negativeness of each experienced change. Positive, negative, and total life change scores are obtained by summing these ratings. (The positive and negative change scores have been found in this and other studies to be essentially uncorrelated with one another.)

In the present study, subjects scoring above and below the median of the positive, negative, and total life change distributions and above and below the median of the SSS were assigned to cells of three separate 2×2 factorial designs. The subjects also completed the Discomfort scale of the Psychological Screening Inventory (Lanyon, 1970), which served as the dependent variable measure of psychological distress.

No significant main or interaction effects were found in analyses of variance involving either the positive or total life change scores. However, a significant main effect for negative life change was found, $F(1, 71) = 4.75$, $p < .05$, with high-change subjects ($M = 10.97$) having higher scores on the Discomfort scale than low scorers ($M = 8.71$). The interaction effect closely approached but did not attain significance. Subsequent Duncan multiple-range tests of cell means showed no significant Discomfort scale differences between high sensation seekers who differed in negative life change. However, among low sensation seekers, high-negative-change subjects ($M = 12.44$) had significantly higher distress scores than did those who had experienced low levels of negative change ($M = 9.00$). The role of sensation seeking as a moderator variable was also indicated in a correlational analysis that disclosed a significant relationship ($r = .35$) between negative life change scores and Discomfort scores in low sensation seekers but no significant relationship ($r = .15$) in high sensation seekers.

The results of this study suggest that life changes are related to psychological distress only if the individual perceives them to be negative, and that the sensation seeking motive influences

Requests for reprints should be sent to Ronald E. Smith, Department of Psychology, NI-25, University of Washington, Seattle, Washington 98195.

the relationship between negative life change and psychological distress.

References

Lanyon, R. I. Development and validation of a psychological screening inventory. *Journal of Consulting and Clinical Psychology*, 1970, 35, 1-24.

Sarason, I. G., & Johnson, J. H. *The Life Experiences Survey: Preliminary findings*. (Tech. Report SCS-LS-001) Seattle, Wash.: Office of Naval Research, 1976.

Zuckerman, M., Kolin, E. A., Price, L., & Zoob, I. Development of a sensation-seeking scale. *Journal of Consulting Psychology*, 1964, 26, 250-260.

Received February 22, 1977 ■

Relation of Patient Attributes to Perceptions of the Treatment Environment

Rudolf H. Moos
Social Ecology Laboratory
Stanford University

Evelyn Bromet
Department of Psychiatry
University of Pittsburgh

The social climate of four residential alcoholism programs was assessed using the Community-Oriented Programs Environment Scale (COPEs). COPEs perceptions were essentially independent of 18 patient background, psychosocial functioning, and alcohol-related characteristics. Considered in conjunction with previous research in other psychiatric and nonpsychiatric settings, the results indicate that perceptions of the social climate of treatment programs are not simply measures of the personal characteristics of the perceivers.

A number of instruments have recently been developed to measure the social-environmental characteristics of psychiatric treatment programs. Several of these techniques assess the treatment environment by asking patients and/or staff about the relevant characteristics of their milieu. In using these "perceived environment" scales, the usual procedure is to average responses from a set of patients and/or staff and to assume that these mean values describe the particular environment being studied.

One problem with this procedure is that there may be substantial differences among individuals in the way in which the "same" treatment environment is perceived. These individual differences have raised the important issue that scales measuring environmental perceptions may reflect background and/or personality characteristics of the perceivers rather than independent attributes of the environment.

The limited empirical evidence that is available indicates that personal characteristics are only minimally related to environmental perceptions (Moos, 1974). The purpose of the present study was to provide more extensive evidence on this issue by relating a wide range of background and intake functioning characteristics to patients' perceptions of the social climate of residential alcoholism programs.

The research population was composed of 326 alcoholic inpatients treated at one of four residential alcoholism programs. Two of the programs were in an urban area (Salvation Army

and a public hospital unit) and were populated by patients who tended to be single, separated, or divorced; residentially mobile; and average or below average in occupational status, income, and education. The other two programs were located in suburban areas and admitted patients who were more often from middle- to upper-middle class backgrounds, who were residentially stable, and who were living with their families. Therefore, we combined patients in the first two programs (Study Group 1; $n = 171$) and those in the latter two programs (Study Group 2; $n = 155$).

Shortly after admission to the programs, patients were administered a detailed background information form, which obtained information about sociodemographic characteristics (nine items), alcohol consumption, behavioral impairment, physical impairment, subjective rating of drinking problem, drinking pattern for the month before entering the program, previous hospitalization for alcoholism, occupational functioning, social functioning, and psychological well-being. There was substantial variability in these sociodemographic and premorbid functioning characteristics due to within-program differences among patients.

Approximately 2-3 weeks after admission, patients were administered the Community-Oriented Programs Environment Scale (COPEs) to evaluate the social climate of their program. The COPEs is composed of 10 subscales: Involvement, Support, Spontaneity, Autonomy, Practical Orientation, Personal Problem Orientation, Anger and Aggression, Order and Organization, Program Clarity, and Staff Control (Moos, 1974). The proportion of subscale variance accounted for by differences among the four programs averaged 20% and ranged from 29.5% for Order and Or-

Requests for reprints and for an extended report of this study should be sent to Rudolf H. Moos, Department of Psychiatry, Stanford University, Stanford, California 94305.

ganization to 9.2% for Program Clarity, indicating that there was substantial subscale variability within each of the programs.

Pearson correlation coefficients between the patient attributes and the 10 COPEs subscales were computed for the two study groups. A series of multiple regression analyses were then performed to assess the overall contributions of the 18 patient attributes to perceptions of the treatment environment. Ten multiple regression analyses (1 for each of the 10 COPEs subscales, which were the dependent variables) were run for each of the two study groups.

Only 7 of the 180 correlation coefficients were .20 or greater in Study Group 1, whereas only 1 of the 180 correlations was .20 or greater in Study Group 2.

Two of the 10 multiple regressions (for Support and Practical Orientation) were statistically significant ($p < .05$) in both study groups; however, none of the 180 possible relationships between the 18 patient background and functioning variables and the 10 COPEs subscales were replicated in the two groups. For example, in Group 1, patients who were better educated and more residentially stable perceived less emphasis on Practical Orientation, whereas patients who were functioning better socially perceived more emphasis on this dimension. These relationships were not replicated in Study Group 2, however, in which Catholic patients and patients who were drinking less and who showed less physical impairment perceived more emphasis on Practical Orientation.

Overall, there was a slight tendency for better educated patients to be somewhat more negative, and for patients with fewer symptoms and less reported drinking severity to be somewhat more positive, in their evaluations of the social climate of their program. However, the correlations were very low and represented less than 5% of the entire correlation matrix. The most parsimonious conclusion to be drawn from these data is that perceptions of treatment environments are only minimally related to patients' background and personality characteristics.

Even though environmental perceptions are independent of a broad array of background characteristics, they are related to an individual's role position in an environment. In general, those responsible for an environment view it more favorably than those not responsible. In addition, staff members of different roles may perceive the same programs differently; for example, nurses and day personnel tend to feel that the professional staff is more supportive than do

aides and evening or night personnel. These findings probably reflect differences in the subenvironments experienced by different individuals.

Environmental perceptions are also related to how well people actually function in an environment. People who see their environments more positively tend to be more satisfied with and perform better in those environments. For example, patients who perceive their programs more positively are more likely to participate in aftercare services (Pratt, Linn, Carmichael, & Webb, 1977), whereas residents who perceive their programs more negatively are more likely to abscond (Chase, 1975, chap. 8).

Although there is little or no relationship between *individual* patient characteristics and their perceptions of the social climates of their treatment programs, the *aggregate* characteristics of patient populations may be related to differences in social climate among treatment settings. However, relationships between aggregate personal and behavioral characteristics and the aggregate social climate cannot be generalized to individuals in any particular setting. Thus, for example, the fact that programs with more disturbed patients have less emphasis on autonomy (Moos, 1974) does not necessarily mean that *within* one program, the more disturbed patients will perceive less emphasis on autonomy than the less disturbed patients.

The present findings demonstrate that *within* programs, personal attributes of patients are only minimally associated with their environmental perceptions. Considered in conjunction with previous research, the results indicate that indices of perceived climate are not simply measures of background or psychosocial characteristics of the perceivers. However, there may be substantial differences in how different individuals perceive the same environment, and it is important to clarify the reasons for these differences. Characteristics of peoples' functioning or performance in an environment might be fruitful to investigate in this regard.

References

- Chase, M. The impact of correctional programs: Absconding. In R. Moos (Ed.), *Evaluating correctional and community settings*. New York: Wiley, 1975.
- Moos, R. *Evaluating treatment environments: A social ecological approach*. New York: Wiley, 1974.
- Pratt, R., Linn, M., Carmichael, J., & Webb, N. The alcoholic's perception of the ward as a predictor of aftercare attendance. *Journal of Clinical Psychology*, 1977, 33, 915-918.

Received March 2, 1977 ■

Effect of Psychotomimetics (LSD and Dextroamphetamine) on the Use of Primary- and Secondary-Process Language

Michael Natale
Department of Psychiatry
Columbia University

Charles C. Dahlberg
William Alanson White Institute
New York City

Joseph Jaffe
Department of Psychiatry, Columbia University and New York State
Psychiatric Institute, New York City

The present study sought to determine the effect of psychotomimetics (LSD and dextroamphetamine) on primary-process and secondary-process language. Five-minute monologues were recorded for four psychoanalytic patients in LSD, dextroamphetamine, and placebo conditions. An intensive research design was adopted, and each subject was treated as an individual experiment: (a) One patient manifested LSD-induced attenuation of secondary-process language; (b) one patient showed LSD-induced increase of primary-process language and an increase of secondary-process language when dextroamphetamine was ingested. The present findings suggest that psychotomimetics do affect ego functioning as expressed in language but that the nature of the effect (inhibition or disinhibition) is not determined by the drug alone.

It has been postulated that psychotomimetics (LSD, Ditrán, psilocybin) have certain consistent effects on verbal behavior. Gottschalk and Gleser (1964) found that LSD, psilocybin, and Ditrán promote disorientation and the use of denial language (subscale of Schizophrenia scale). Fink (1974) claims that psychotomimetics, such as dextroamphetamine, Ditrán, and LSD cause a decrease in "defensive language" (denial and disorientation). The contradictory nature of these findings may be the result of the fact that Gottschalk and Gleser's results were obtained from normals, and Fink's data were obtained from psychotic depressed patients who had received electric shock.

In light of the above-described "state of knowledge," the purpose of the present study was to verify the effects of psychotomimetics on

ego functioning as expressed in language use. "Pre-drug" and "post-drug" 5-minute monologues were recorded for four psychoanalytic patients (Jaffe, Dahlberg, Luria, & Chorosh, 1973) in LSD (50-100 mg), dextroamphetamine (15 mg), and placebo conditions. Each patient participated in 6-9 sessions for each drug condition, and an "intensive research design" was adopted, with each patient being treated as a separate experiment. Each monologue was transcribed, key-punched, and analyzed by Martindale's (1973) "count" program, which tabulates the percent frequency of words that belong to primary-process or secondary-process language categories. The results were as follows: (a) One patient manifested a significant attention of secondary-process language when under the influence of LSD, $t(7) = 2.48$, $p < .05$, two-tailed, with no significant placebo effect. (b) One patient showed a significant increase of secondary-process vocabulary when under the influence of dextroamphetamine, $t(8) = 3.07$, $p < .05$, two-tailed, and a significant increase in primary-process language after having taken LSD, $t(5) = 2.58$, $p < .05$, two-tailed, with no significant placebo effect. (c) Two patients manifested no drug or placebo effects. The present and previous findings suggest that psychotomimetics do affect ego functioning, as expressed in language, but that the nature of the effect (inhibitory or disinhibitory) is not determined by the drug alone.

This research was aided in part by National Institute of Mental Health Grant MH 13579, which supported the first author as a postdoctoral research fellow in the Interdisciplinary Research Training Program, Department of Psychiatry, Columbia University.

The authors wish to express thanks to the William Alanson White Institute where the project was originally conducted.

Requests for reprints and for an extended report of this study should be sent to Michael Natale, who is now at the Department of Psychology, Whitley Psychology Laboratories, Franklin and Marshall College, Lancaster, Pennsylvania 17604.

References

- Fink, M. Brain behavior, verbal behavior, and psychotherapy. *Comprehensive Psychiatry*, 1974, 15, 257-266.
- Gottschalk, L. A., & Gleser, G. C. Distinguishing characteristics of verbal communication with schizophrenics. In D. M. Rioch & E. A. Weinstein (Eds.), *Disorders of communication*. Baltimore, Md.: Williams & Wilkins, 1964.
- Jaffe, J., Dahlberg, C. C., Luria, J., & Chorosh, J. Effects of LSD-25 and dextroamphetamine on speech rhythms in psychotherapy dialogues. *Biological Psychiatry*, 1973, 6, 93-96.
- Martindale, C. Count: A PL/9 program for content analysis of natural language. *Behavioral Science*, 1973, 18, 148.

Received March 3, 1977 ■

Short Forms of the MMPI with Back Pain Patients

Judith Turner
University of California, Los Angeles

Charles McCreary
Department of Psychiatry
University of California, Los Angeles
School of Medicine

This study compared the Faschingbauer abbreviated Minnesota Multiphasic Personality Inventory (FAM) and the Midi-Mult on a sample of 176 back pain outpatients. Correlations with the standard MMPI scales ranged from .67 to .93 ($M = .84$) on the FAM and from .71 to .89 ($M = .80$) on the Midi-Mult. The FAM showed higher agreement with MMPI code types than did the Midi-Mult. Finally, the three versions were compared to independent physician ratings of amount of functional component to patients' pain, and all three forms discriminated between "functional" and "organic" patient groups. The results provide tentative evidence that abbreviated MMPIs are useful measures of personality in this population.

A number of short forms of the Minnesota Multiphasic Personality Inventory (MMPI) have been constructed in attempts to produce an instrument that gives the same information but with much less time required of the respondent. The present investigation addressed the issue of the utility of two of these short forms—the Midi-Mult (Dean, 1972) and the Faschingbauer abbreviated MMPI (FAM) (Faschingbauer, 1974)—in a university hospital outpatient orthopedic clinic.

The standard form of the MMPI was administered to 186 male and female outpatients (M age = 44 years) who were seen at the Back Clinic. Each answer sheet was scored for the standard MMPI, Midi-Mult, and FAM scales except that 10 MMPIs were excluded because more than 30 items had been left unanswered. Ratings by the examining physician as to the amount of functional component to the patient's pain were recorded for 64 patients at the time of the patient's first visit to the Back Clinic and were made without knowledge of the patient's MMPI performance.

MMPI profiles were called invalid according to these rules: $F > 16$ and $F - K > 11$ raw score points. These rules were applied separately for the three versions. Two percent of the MMPIs were found to be invalid, whereas 5% of the FAM profiles were invalid, including 75% of the invalid MMPI profiles. None of the Midi-Mult profiles were invalid, showing a 100% error rate in iden-

tifying invalid standard MMPI profiles. The four standard MMPI profiles found to be invalid were excluded from further analysis of data.

Midi-Mult scale scores correlated with corresponding standard MMPI scale scores from .69 to .89 ($M = .80$). Correlations between FAM and full MMPI scales ranged from .67 to .93 ($M = .85$). The FAM was more highly correlated with the full MMPI than was the Midi-Mult on all clinical scales except Hypochondriasis, whereas the Midi-Mult was more highly correlated with the standard form on F and K .

Next, code-type agreement between the standard MMPI and the two short forms was examined. On a two-point code type comparison, the FAM was found to have 34% agreement with the long version in the same order and 43% agreement regardless of order. The Midi-Mult showed 22% agreement (same order) and 32% (any order). Less agreement resulted when three-point codes were compared (FAM: 14% same order, 31% any order; Midi-Mult: 9% same, 20% any).

Finally, comparisons were made between standard MMPI, FAM, and Midi-Mult profiles and independent physician ratings of the amount of functional component to patients' pain. Patients judged by physicians to have a very large functional component ($n = 39$) had standard MMPI scores that were significantly higher than those of patients judged to have little functional component ($n = 25$) on the K , Hypochondriasis, Hysteria, Schizophrenia, and Social Introversion scales ($p < .05$) and Psychopathic Deviate scale ($p < .01$). A similar pattern of differences was shown by the FAM, except that there were no significant differences between "functionals" and "organics" on the K and Psychopathic Deviate scales, whereas

Requests for reprints and for an extended report of this study should be sent to Charles McCreary, Department of Psychiatry, University of California, Los Angeles, School of Medicine, 760 Westwood Plaza, Los Angeles, California 90024.

the "functional" patients scored significantly higher on the Depression and Psychasthenia scales ($p < .05$). The Midi-Mult also showed a pattern of differences similar to the MMPI, except for no significant difference on the K scale and significant differences on Depression, Paranoia, and Psychasthenia ($p < .05$).

That both short forms were found to discriminate between "functional" and "organic" back pain patients is encouraging evidence that these measures may be useful with this population. Since the FAM, in particular, was shown to provide a

very good estimate of standard MMPI scores, further attention seems warranted in the direction of its use in such a setting.

References

- Dean, E. A lengthened Mini: The Midi-Mult. *Journal of Clinical Psychology*, 1972, 28, 68-71.
- Faschingbauer, T. R. A 166-item written short form of the group MMPI: The FAM. *Journal of Consulting and Clinical Psychology*, 1974, 42, 645-655.

Received March 7, 1977 ■

Practicing A and B Psychotherapists' Responses to Schizoid and Neurotic Patient Prototypes

William B. Goodwin, Donald M. Quinlan, and Jesse D. Geller
Yale University and Yale School of Medicine

Eighty-three practicing male psychotherapists completed the A-B scale, responded to recordings of schizoid and neurotic patient prototypes, and rated their subjective reactions to each type. Multivariate analysis revealed a significant ($p < .013$) overall A-B Type \times Patient Type interaction. Although liking and ease of responding were higher in therapist-patient dyads, which the literature suggests are effective (A-schizoid, B-neurotic), felt compatibility and desire to work with the patient were higher in the opposite ("mismatched") dyads. These results suggest that subjective reactions underlying the A-B interaction effect are complex and that therapists respond in a differentiated, not global, fashion.

The A-B therapist "type" distinction originated by Whitehorn and Betz (cf. Betz, 1967) is an intriguing but conceptually elusive variable associated with a differential therapeutic aptitude. Based on items from the Strong Vocational Interest Blank reflecting activities mostly technical, mechanical, or manual in nature, the A-B scale differentiated therapists more effective with schizophrenics (As) from those less effective (Bs). McNair, Callahan, and Lorr (1962) later found Bs to be more effective with neurotic outpatients, giving rise to the "A-B interaction hypothesis" that As and Bs are differentially effective with schizophrenic and neurotic patients, respectively. Owing to the lack of clinical studies with the requisite factorial design, the interaction hypothesis has not yet been fully evaluated, although one such study appears to provide support (Berzins, Ross, & Friedman, 1972).

Experimental studies using schizoid and neurotic patient stimulus materials have given rise to contradictory findings regarding therapists' subjective reactions, however. Although Kemp (1966), for example, found a "paradoxical discomfort" on the part of As with schizoid "patients" and Bs with neurotic patients, Berzins and Seidman (1968) failed to replicate. As a whole, such analogue studies have lacked realism in numerous respects: in the nature of the patient stimuli, in the mode of response, in the social context, in the physical setting, and in the

sample of subjects used as "therapists." Further, the impact of therapist experience level on the A-B variable has not been evaluated. Our primary aim was to extend findings of subjective reactions of A and B undergraduate "quasi-therapists" to a sample of practicing A and B psychotherapists, using the experimental analogue approach.

Eighty-three out of 136 practicing male psychotherapists with at least 1 year of experience completed the Schiffman, Carson, and Falkenberg (Note 1) 23-item version of the A-B scale. The 20 highest scoring (A) and 20 lowest scoring (B) therapists were presented realistic audio recordings of avoidant of others (schizoid) and turning against self (neurotic depressive) patient prototypes used by Berzins and Seidman (1968), listened to five 1-minute segments of each patient in the privacy of their own offices, and spoke their responses—which could include interrupting the patient or silence—into a cassette recorder. Order of presentation of prototypes was counterbalanced across therapists. After each 5-minute segment, therapists completed a six-item questionnaire assessing subjective reactions to the patient.

Results

A four-factorial multivariate analysis of variance (Therapist Type \times Experience Level \times Patient Type \times Order) yielded a significant ($p = .013$) Therapist Type \times Patient Type interaction. The standardized discriminant function coefficients weighted the six components as follows: (a) desire to work with the patient, -1.165 ; (b) ease of responding, $.832$; (c) lik-

Requests for reprints and for an extended report of this article should be sent to Donald M. Quinlan, Memorial Unit 10-East, Yale-New Haven Hospital, 789 Howard Avenue, New Haven, Connecticut 06504.

ing, .816; (d) felt compatibility, $-.328$; (e) felt comfort, $.171$; and (f) satisfaction with spoken responses, $-.137$. Two univariate analyses of variance revealed trends for reversed Therapist Type \times Patient Type interactions: Higher scores were obtained for "mismatched" (B schizoid, A neurotic) therapists on desire to work with the patient ($p = .055$) and felt compatibility ($p = .09$). No significant main effects were found nor were there any significant interactions with experience or order of presentation.

Discussion

Broadly summarized, although these therapists felt it easier to respond to and like the "matched" patient better, they felt more compatible with and preferred to work with the mismatched patient. These results offer support for the hypothesis that A and B psychotherapists can be differentiated by their subjective reactions to schizoid and neurotic patients. The direction and meaning of these differences is complex.

Previous investigators have hypothesized that failures of therapeutic communication can result from resemblances between patient defense mechanisms and therapist coping style, and that there are certain personological similarities between Bs and schizoids and possibly between As and neurotic depressives (e.g., Berzins, Seidman, & Welch, 1970). Taken together, these considerations suggest that (a) therapists may equate similarity with compatibility and therefore prefer to work with patients because they perceive them as similar, hence compatible or (b) if therapists are aware of their limitations with mismatched patients, their preferences may reflect a desire to grapple with and overcome their "blind spots" by means of massed practice. There is some suggestion (Goodwin, in press) that inexperienced B therapists do see a higher proportion of schizoid patients than inexperienced As. If this is a replicable finding, the possibility that B (and/or A) therapists increase their clinical effectiveness with mismatched patients as they gain experience needs empirical study. A third explanation for the present findings is that some degree of felt incompatibility is more favorable to therapeutic outcome.

If all of the subjective reactions—liking, ease of responding, felt compatibility, and so on—

showed the same pattern of higher scores for matched dyads, we might have speculated that there is a simple underlying affective dimension describable as "good vibrations." However, the different signs of the discriminant function coefficients and the contrasting patterns of means indicate that the subjective *alternative* reactions underlying the A-B interaction effect are complex, and that more broadly, therapists' reactions to patients are differentiated rather than unidimensional. Finally, how therapists use such feelings—"countertransference" or otherwise—as a data base from which to formulate therapeutic responses is a question that deserves empirical study.

Reference Note

1. Schiffman, H., Carson, R. C., & Falkenberg, P. A psychometric analysis of the Whitehorn-Betz A-B scale. Unpublished manuscript, Duke University, 1967.

References

- Berzins, J. I., Ross, W. F., & Friedman, W. H. A-B therapist distinction, patient diagnosis, and outcome of brief psychotherapy in a college clinic. *Journal of Consulting and Clinical Psychology*, 1972, 38, 231-237.
- Berzins, J. I., & Seidman, E. Subjective reactions of A and B quasi-therapists to schizoid and neurotic communications. *Journal of Consulting and Clinical Psychology*, 1968, 32, 342-347.
- Berzins, J. I., Seidman, E., & Welch, R. D. A-B therapist "types" and responses to patient-communicated hostility: An analogue study. *Journal of Consulting and Clinical Psychology*, 1970, 34, 27-32.
- Betz, B. J. Studies of the therapist's role in the treatment of the schizophrenic patient. *American Journal of Psychiatry*, 1967, 123, 963-971.
- Goodwin, W. B. Caseloads of A and B psychotherapists: Paradoxical preference? *Psychotherapy: Theory, Research and Practice*, in press.
- Kemp, D. E. Correlates of the Whitehorn-Betz A-B scale in a quasi-therapeutic situation. *Journal of Consulting Psychology*, 1966, 30, 509-516.
- McNair, D. M., Callahan, D. M., & Lorr, M. Therapist "type" and patient response to psychotherapy. *Journal of Consulting Psychology*, 1962, 26, 425-429.

Received March 24, 1977 ■

Comparisons of Rating Scales of Child Psychopathology in Clinic and Nonclinic Samples

Susan B. Campbell

McGill University and the Montreal Children's Hospital

Yvonne Steinert

Jewish General Hospital
Montreal, Canada

Mothers of 45 control and 35 clinic children completed two factor-analytically derived rating scales of child psychopathology, the Behavior Problem Checklist and the Parent Questionnaire. Intercorrelations among factors indicated some factor overlap and some differences between apparently similar factors. Conduct problem ratings on the Behavior Problem Checklist covaried with ratings of conduct problems, hyperactivity, and learning problems on the Parent Questionnaire. Patterns of correlations suggested that mothers of nonreferred children rated pathology *per se*, whereas mothers of referred children rated behaviors that fell into internalizing and externalizing clusters. Teachers asked to rate 22 referred and 22 control children on the Behavior Problem Checklist and the Teacher Rating Scale showed more consistency in the behaviors rated. Although similar factor labels are used from scale to scale, they reflect somewhat different behaviors and are influenced by the child's clinical status, the rater's past experience with children, and specific scale characteristics.

The Behavior Problem Checklist (Quay & Peterson, Note 1), the Parent Questionnaire (Conners, 1970), and the Teacher Rating Scale (Conners, 1969) are factor-analytically derived instruments based on common descriptors of child psychopathology. They are widely used in epidemiological, diagnostic-descriptive, and outcome studies with child populations (Quay, 1972). However, despite overlap and variation in both item content and factor labels, it is unclear how the factor scores covary.

Content analysis suggests that different behaviors are covered by the same factor label and similar behaviors by different labels from scale to scale. Thus, all include a conduct problem factor reflecting discipline problems and defiance of authority. While one scale also includes hyperactive behavior under this rubric, the others isolate it as a separate factor. Likewise, all include a neurotic factor with items about fears and worries but differing in emphasis on social withdrawal and low self-esteem. Thus, the present study provides data on the relationships among those factors with overlap in content and/or label. Several factors on

the Parent Questionnaire that load on only a few items and are infrequently rated were excluded, as was the socialized delinquency factor on the Behavior Problem Checklist, which has little relevance with younger age groups.

Forty-five control and 35 clinic children (M ages = 7.6 and 7.5) took part in this study. Non-clinic subjects were at grade level and had never received psychological help. Clinic subjects were perceived as inattentive and poor school achievers with a variety of additional behavior problems. There were five girls in each group. All children were of at least average IQ, from lower middle to middle class, and in kindergarten through Grade 4 in school. Mothers completed both the Parent Questionnaire and the Behavior Problem Checklist by rating a series of statements as no problem, a mild problem, or a severe problem. In addition, six classroom teachers completed the Behavior Problem Checklist and the Teacher Rating Scale on a similar and partially overlapping sample of 22 clinic and control boys (M ages = 8.3 and 8.5).

Factor scores on the two scales completed by mothers were correlated for clinical and control groups separately. Similar analyses were carried out on teacher ratings. Data are summarized in Table 1. Ratings of clinic children by mothers suggest consistency between scales in that conduct problems, learning problems, and hyperactivity all intercorrelated, indicating a cluster of externalizing behaviors. It appears from the pattern of relationships that conduct problems as measured by the Behavior Problem Checklist include discipline

This research was supported by Grant MA-4505 from the Medical Research Council of Canada to the first author.

Requests for reprints should be sent to Susan Campbell, who is now at the Department of Psychology, Clinical Psychology Center, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

Table 1
Correlations Between Behavior Problem Checklist and Conners' Parent and Teacher
Rating Scale Factor Scores

Scale	Peterson-Quay behavior problem checklist					
	Control			Clinic		
	Conduct problem	Personality problem	Inadequacy-immaturity	Conduct problem	Personality problem	Inadequacy-immaturity
Parent Questionnaire						
Conduct Problem	.498***	.283	.310*	.747***	.266	.262
Anxiety	.354*	.191	.387**	.214	.562**	.587**
Hyperactivity	.533***	.479***	.337*	.824***	.257	.221
Learning Problem	.421**	.253	.560***	.635**	.279	.268
Psychosomatic	.252	.137	.264	.177	.134	.340*
Teacher Rating Scale						
Conduct Problem	.155	-.023	.506*	.536**	.135	.448*
Inattentive-Passive	.537**	.162	.580**	.416*	.245	.648***
Tension-Anxiety	-.115	.372	.248	-.349	.456*	.122
Hyperactivity	.921***	.190	.215	.569**	-.124	.567**

Note. For the Parent Questionnaire, $n_s = 45$ and 35 for control and clinic samples, respectively. For the Teacher Rating Scale, $n = 22$ for both the control and clinic samples.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

problems, attentional difficulties, and dislike of school. Similarly, immaturity, anxiety, and personality problems are interrelated, indicating a cluster of internalizing symptoms, with the anxiety factor including behaviors measured by both the immaturity and personality problem factors. Psychosomatic complaints are relatively independent of other factors. Although control mothers' ratings form similar patterns, additional correlations between internalizing and externalizing factors, for example, conduct problem and anxiety, suggest that they are rating a more heterogeneous array of pathology rather than behaviors that cluster into "neurotic" or "acting out" forms of pathology. Thus, the distinction between internalizing and externalizing symptoms appears more evident in the context of either recognized pathology or a larger number of deviant behaviors.

Teacher ratings form similar patterns of correlations for both groups. Thus, inattentive and hyperactive behaviors show strong relations to conduct problems. However, conduct problem ratings appear to reflect somewhat different behaviors in control and clinic groups. Teachers appear to be rating primarily fidgety, impulsive, and restless behavior in the control group, whereas conduct problem ratings in the clinic group appear to also reflect discipline problems. Moreover, teachers, who have experience with a large number of children, appear to group internalizing and externalizing behaviors intuitively whether they are rating clinic

or control groups, something not evident in the ratings of control mothers.

These data indicate that behaviors considered indicative of child psychopathology differ in patterning as a function of the child's clinical status and with the experiences of the rater. In addition, there is evidence that these scales measure both overlapping and distinctive aspects of behavior problems. This appears to be the case even for factors with the same name. Thus, studies that define samples on the basis of high scores on factors such as conduct problem or anxiety need to consider differences among scales as well as characteristics of raters and subjects.

Reference Note

1. Quay, H. C., & Peterson, D. K. *Manual for the Behavior Problem Checklist*. Unpublished manuscript, University of Illinois, 1967. (Mimeo)

References

- Conners, C. K. A teacher rating scale for use in drug studies with children. *American Journal of Psychiatry*, 1969, 126, 152-154.
- Conners, C. K. Symptom patterns in hyperkinetic, neurotic, and normal children. *Child Development*, 1970, 41, 667-682.
- Quay, H. C. Patterns of aggression, withdrawal, and immaturity. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood*. New York: Wiley, 1972.

Received March 28, 1977 ■

Buss-Durkee Assessment and Validation with Violent Versus Nonviolent Chronic Alcohol Abusers

Gisele J. Renson, John E. Adams, and Jared R. Tinklenberg

Department of Psychiatry and Behavioral Sciences, Stanford School of Medicine,
and Veterans Administration Hospital, Palo Alto, California

Twenty-six chronic alcohol abusers who had been violent while intoxicated and 25 nonviolent alcohol abusers were administered the Buss-Durkee Inventory. All subjects were Caucasian men with a reported daily intake of ethanol of $227 \text{ ml} \pm 89 \text{ ml}$ for at least the last 5 years. Violence was documented by police records and by patient and family reports. Violent drinkers scored significantly higher than control subjects on the inventory Total Hostility score and on subscales measuring Assault, Irritability, Verbal Hostility, Indirect Hostility, and Resentment.

The earliest hostility inventories developed during the late 1950s, with few exceptions (Schulz, 1954; Zaks & Walters, 1959), consolidated various aggressive behaviors into a single omnibus score (Buss, 1961). Buss and Durkee (1957) developed subclassifications of overt and covert hostility. Normative data for this inventory have been reported by Buss (1961) for adult normal and psychiatric populations and by Morrison, Chaffin, and Chase (1975) for normal adolescents. The present study attempted to validate the Buss-Durkee as an inventory of hostility and to provide normative data for a population of violent and nonviolent alcohol abusers. We hypothesized that the Buss-Durkee Inventory would discriminate between these two groups and that overt hostility would be higher in the scores of alcoholics who became violent while intoxicated.

The experimental subjects were 26 Caucasian men ($M \text{ age} = 37 \pm 10$), who were seen as outpatients at the Alcohol and Violence Clinic of the Stanford University Department of Psychiatry and Behavioral Sciences. All of these subjects had demonstrated alcohol abuse, and virtually all (25 of 26) had demonstrated violent behavior only

while intoxicated. A control group of 25 Caucasian men ($M \text{ age} = 43 \pm 9$) was selected from the Stanford Inpatient Service for alcohol abusers at the Palo Alto Veterans Administration Hospital. None of the control subjects had demonstrated violence while intoxicated.

All subjects were screened by one of the authors and at least one other member of the Stanford psychiatric staff to exclude patients with neurological disorders or history of other forms of drug abuse. Experimental and control groups were matched for age, education, occupational level, and alcohol consumption. The majority of subjects in both groups were unskilled or semiskilled workers (65% of the experimental and 72% of the control subjects), according to the Hollingshead (Note 1) classification. All subjects had had a daily ethanol consumption of $227 \text{ ml} \pm 89 \text{ ml}$ for the past 5 years or more, and most had begun drinking heavily in their late teens (16 ± 3 years).

Experimental subjects had police-documented histories of violence while intoxicated. Twenty-four experimental subjects had documented physical assaults against others. Both of the remaining experimental subjects had histories of verbal hostility toward their families. In addition, one had threatened his wife with a deadly weapon and the other had a police record for destruction of property.

Within the first 2 weeks of admission to treatment, each subject completed and signed the Buss-Durkee Inventory under the staff psychologist's supervision. All subjects were assured complete confidentiality.

The Buss-Durkee Inventory (Buss & Durkee, 1957) is a self-rating scale of 75 true-false items based on the rationale that hostility can usefully be divided into seven subgroups or scales: (a)

This research was supported by National Institute for Alcohol Abuse and Alcoholism Grant AA 00498, the Commonwealth Foundation, and the Veterans Administration.

We thank C. P. C. Rosenbaum for his assistance.

John E. Adams is now at the Department of Psychiatry, University of Florida Medical Center.

Requests for reprints and for an extended report of this study should be sent to Jared R. Tinklenberg, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California 94305.

Table 1
Buss-Durkee Inventory Scores of Violent
(Experimental) and Nonviolent (Control)
Alcohol Abusers

Scale	Experimental ^a		Control ^b		<i>t</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Total Hostility	36.58	9.07	28.64	9.40	3.07**
Assault	6.15	2.33	4.40	1.76	3.02**
Indirect					
Hostility	5.54	1.79	4.36	2.00	2.22*
Irritability	6.88	2.45	4.92	2.56	2.79**
Negativism	2.35	1.23	2.48	1.60	.33
Resentment	3.32	1.86	2.60	2.00	1.33*
Suspicion	3.88	1.77	3.28	2.30	1.05
Verbal					
Hostility	8.58	2.76	6.36	2.58	2.96**
Guilt	5.12	2.53	5.58	2.62	.64

^a *n* = 26.

^b *n* = 25.

* *p* < .05.

** *p* < .01.

Assault or direct physical violence against persons; (b) Indirect Hostility either against persons, through gossip or practical jokes, or against objects, such as slamming doors or breaking things; (c) Irritability or explosiveness and exasperation at the slightest stimulus; (d) Negativism as either active rebellion or as passive compliance to rules and authority figures; (e) Resentment, anger, jealousy, and/or hate of others due to real or imaginary mistreatment; (f) Suspicion, varying from distrust to projections of hostility onto others and the belief that these others are derogatory and harmful; and (g) Verbal Hostility in style or content. These seven scales yield a Total Hostility score. There is also a Guilt scale, which is independent of the above.

Table 1 shows the scores obtained by the control and experimental groups. The Total Hostility score reveals that violent alcohol abusers scored significantly higher ($p < .01$) than the nonviolent group. A more refined presentation of the kinds of hostility experienced by the subjects is reflected in the subscale scores. Violent alcoholics scored significantly higher ($p < .01$) on the scales measuring Assault, Irritability, and Verbal Hostility, and also ($p < .05$) on Indirect Hostility and Resentment. The two groups did not differ significantly on Negativism, Suspiciousness, or on the Independent Guilt scale.

The data were analyzed for any relationship between age, individual scores, and/or Total Hostility score. Both groups were divided into younger (up to 40 years) and older (40 years and above) subjects. A linear regression procedure showed no relationship between age and any scores for the control subjects. For the experimental subjects, an age effect was shown for the Suspicion scale only, which tended to decrease with age ($p < .05$). The five most verbally aggressive subjects were compared with the five least verbally aggressive subjects in each group to assess whether subjects with higher verbal aggressiveness scores tended to have lower assaultiveness scores. This was not found to be the case; subjects with higher verbal aggressiveness scores did not have lower assaultiveness scores. All subjects had high scores on the Assault scale.

Our findings show a positive correlation between violence while intoxicated and Buss-Durkee Inventory scores on Total Hostility as well as five of seven hostility subscales among two groups of alcohol abusers. This confirms the validity of the Buss-Durkee Inventory as a measure of hostile feelings and aggressive behavior in this sample and suggests that the Buss-Durkee Inventory may be useful in assessing potential for violence in alcohol abusers.

Reference Note

1. Hollingshead, A. B. *Two-factor index of social position*. Unpublished manuscript, Yale University, 1957.

References

- Buss, A. H. *The psychology of aggression*. New York: Wiley, 1961.
- Buss, A. H., & Durkee, A. An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology*, 1957, 21, 343-349.
- Morrison, S. D., Chaffin, S., & Chase, T. V. Aggression in adolescents: Use of the Buss-Durkee Inventory. *Southern Medical Journal*, 1975, 68, 431-536.
- Schultz, S. D. *A differentiation of several forms of hostility by scales empirically constructed from significant items of the Minnesota Multiphasic Personality Inventory*. Unpublished doctoral dissertation, Pennsylvania State College, 1954.
- Zaks, M. S., & Walters, R. H. First steps on the construction of a scale for the measurement of aggression. *Journal of Psychology*, 1959, 47, 199-208.

Received April 14, 1977 ■

On the Word Preferences of Suicidal Versus Nonsuicidal College Students

Steven Thurber and David P. Torbet
Boise State University

A word preference format was used to investigate reactions to verbal stimuli of suicidal and nonsuicidal persons. Six words with either aggressive or submissive denotative meanings significantly differentiated the two groups. In addition, the word *suicide* was selected at a higher frequency level by suicidal individuals when compared to their nonsuicidal counterparts.

Jung (1918) and Luria (1932) were notable among several early clinicians who investigated the premise that affective reactions to verbal stimuli might yield information as to personality characteristics and motivational states. The traditional approach was to base clinical inferences on the denotative meaning of words eliciting high levels of emotionality as judged by physiologic indicators (e.g., skin resistance; changes in breathing). More recently, affective arousal (or what has since been termed *reinforcement value*) has been assessed by ratings on a like-dislike continuum (Rychlak, 1966) or by noting word preferences in a series of paired associates (Torbet, Note 1).

The present study was designed to see whether suicidal individuals could be distinguished from a nonsuicidal group on the basis of word preferences. The participants were 21 persons referred to a university counseling center following documented instances of attempted suicide (2 of these individuals were later successful in suicide attempts). They were compared to students from general psychology classes controlling for age ($M = 25.9$) and sex (13 females; 8 males). The subjects selected the most preferred word for each of 233 word pairs. Randomly distributed among these items were six words expected to elicit differential reactions from the *a priori* groups. They included the words *murder*, *kill*, *attack*, *submit*, *suicide*, and *martyr*. Together, they formed 30 word pairs, and in accord with circular unfolding scaling procedures they constituted an undimensional metric scale or what was termed an *aggressive-*

submissive dimension (Cooper, 1971). Following the designation of preferences across all items, the frequencies within this dimension were tallied, with each word having a possible score of 5.

Wilk's lambda criterion indicated that the equality of mean vectors for the two groups was untenable, $\chi^2(6) = 17.96$, $p < .006$. Univariate analyses indicated that the suicidal group showed a significantly higher preference for suicide ($M = 3.76$ vs. $M = 1.90$), $t(20) = 4.47$, $p < .001$, whereas the nonsuicidal subjects significantly preferred attack ($M = 3.38$ vs. $M = 2.62$), $t(20) = 2.61$, $p < .01$.

A linear discriminant analysis yielded a canonical correlation of .62. The primary contributor to the discriminant function was suicide, with a standardized coefficient of 1.29 followed in order of magnitude by submit (.56), martyr (.52), kill (.48), murder (.29), and attack (.02). Evidence for the discriminating ability of the word scale is shown in the finding that 34 of the 42 participants (81%) were correctly classified into suicidal and nonsuicidal categories on the basis of discriminant scores.

The results suggest the potential clinical utility of a word preference technique in the assessment of suicidal probabilities. The denotative meanings of the stronger discriminating variables support the notion that attempted suicide may represent a surrendering to hostile impulses by turning them toward oneself in an intropunitive manner (Kisker, 1964). Finally, the data suggest that the word *suicide* has reinforcement value for those with suicidal inclinations.

Reference Note

1. Torbet, D. P. *Manual for the Word Preference Personality Inventory*. Unpublished manuscript, Boise State University, 1976.

Requests for reprints and for an extended report of this study should be sent to David P. Torbet, Director, Counseling and Testing Center, Boise State University, Boise, Idaho 83725.

References

- Cooper, T. T. Circular unfolding theory. (Doctoral dissertation, University of Michigan, 1971). *Dissertation Abstracts International*, 1971, 32, 3888A. (University Microfilms No. 72-4847, 77)
- Jung, D. G. *Studies in word association*. London: Heinemann, 1918.
- Kisker, G. W. *The disorganized personality*. New York: McGraw-Hill, 1964.
- Luria, A. R. *The nature of human conflicts, or emotions, conflict and will*. New York: Liveright, 1932.
- Rychlak, J. F. Reinforcement value: A suggested idiographic, intensity dimension of meaningfulness for the personality theorist. *Journal of Personality*, 1966, 34, 311-335.

Received May 9, 1977 ■

Role of Locus of Control in Frustration-produced Aggression

Kiran Bhatia and Sanford Golin
University of Pittsburgh

This experiment tested the hypothesis that external locus of control subjects would exhibit greater frustration-produced aggression than internal locus of control subjects. Incorrect responses by a frustrating or nonfrustrating confederate were punished by electric shock. Analysis of the shock data indicated increased aggression after frustration by externals and less aggression by internals. The results support the hypothesis and indicate that aggression is cognitively regulated by a personality-related belief in uncontrollability; the less the believed control, the greater the aggression.

This experiment was concerned with the cognitive control of frustration-produced aggression. According to one view (e.g., Geen, 1972), frustration results in aggression when it increases arousal in the presence of cues associated with aggression. Recent reports indicate that the extent to which frustrating events are perceived as aversive and thereby increase arousal is a function of the extent to which such events are perceived as controllable (Donnerstein & Wilson, 1976; Glass & Singer, 1972). Perception of controllability is related to the characteristics of the particular situation in which frustration occurs. Such perception, however, may have personality as well as situational determinants. Specifically, one's locus of control (see Phares, 1976), that is, transsituational beliefs about one's ability to exercise control over outcomes, may influence the extent to which frustration results in aggression. Those who generally do not believe they can control outcomes (external locus of control) are expected to exhibit greater aggression in response to frustration than those who generally believe they can control outcomes (internal locus of control). The present experiment tested this hypothesis.

Forty-eight internal locus of control and 48 external locus of control subjects were equally divided into frustration and nonfrustration experimental subgroups. Using a standard procedure to measure aggression, subjects taught a confederate a task by punishing incorrect responses

with electric shock, which was ostensibly administered to the confederate. Shock durations and intensities were recorded as measures of aggression. Each subject was motivated to have the confederate learn the task as quickly as possible; the confederate learned relatively poorly in the frustration condition and relatively well in the nonfrustration condition.

An analysis of variance of shock duration data showed a Locus of Control \times Frustration interaction, $F(1, 88) = 5.86, p < .05$. Scheffé's test ($p < .05$) showed that the mean shock duration of frustrated externals ($M = .65$ sec) was greater than the means of nonfrustrated externals ($M = .49$ sec) and frustrated internals ($M = .50$ sec). Frustrated internals showed less aggression than nonfrustrated internals ($M = .58$ sec), but this difference was not significant. As predicted, therefore, frustration resulted in increased aggression for externals but not for internals.

Analysis of the shock intensity data showed the mean intensity of frustrated internals ($M = 2.97$) to be less ($p < .05$) than that of nonfrustrated internals ($M = 4.37$). The mean intensity of frustrated externals ($M = 5.10$) was greater than that of nonfrustrated externals ($M = 4.32$), as predicted, but this difference was not significant.

The prediction that externals would exhibit greater aggression in response to frustration was based on the view that aggression would be regulated by a generalized belief in uncontrollability. If, however, locus of control were considered to be a measure of a need to control, a motive said to be characteristic of internals (see Phares, 1976), then frustration should have been more aversive and arousing for internals than for externals. A prediction of greater aggression in response to frustration by internals follows from this perspective. Internals, how-

This article is based on a dissertation submitted by the first author in partial fulfillment for the PhD degree at the University of Pittsburgh.

Requests for reprints and for an extended report of this study should be sent to Sanford Golin, Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania 15270.

ever, showed reduced rather than increased aggression in response to frustration. Hence, the present results were not in accord with this motivational interpretation of locus of control.

In summary, the results showed that generalized expectancies about one's ability to control outcomes can influence aggression in response to frustration in a manner similar to that previously reported for situationally induced expectancies about controllability (Donnerstein & Wilson, 1976): The less the believed control, the greater the aggression.

References

- Donnerstein, E., & Wilson, D. W. Effects in noise and perceived control on ongoing and subsequent aggressive behavior. *Journal of Personality and Social Psychology*, 1976, 34, 774-781.
- Geen, R. G. *Aggression*. Morristown, N.J.: General Learning Press, 1972.
- Glass, D., & Singer, J. *Urban stress*. New York: Academic Press, 1972.
- Phares, E. J. *Locus of control in personality*. Morristown, N.J.: General Learning Press, 1976.

Received May 9, 1977 ■

Personality and Social Factors in Adolescent Marijuana Use: A Path-Analytic Study

Paul M. Kohn

York University, Downsview, Canada
and Addiction Research Foundation
Toronto, Canada

Helen M. Annis

Addiction Research Foundation
Toronto, Canada

Measures of the following variables were obtained from 193 high school students: marijuana use; attitude toward marijuana use; the peer-acceptance and symbolic-protest functions of marijuana use; sociopolitical outlook; and internal sensation seeking. Path analysis supported a model that assumes the following: (a) Only attitude affects use directly. (b) Both the peer-acceptance and symbolic-protest functions, as well as sociopolitical outlook, and internal sensation seeking influence attitude directly. (c) Sociopolitical outlook affects both functions directly. Support for the model came from very close correspondence between the observed and predicted correlations and the low, nonsignificant value of the overidentification test statistic.

This study evaluates a multivariate model of youthful marijuana use. The model assumes that use or nonuse reflects a person's attitude toward marijuana use, that is, his/her evaluative reaction to such behavior (cf. Fishbein, 1967a, 1967b). Attitude, in turn, depends on the perceived functions of marijuana use. Our model includes two functions: the symbolic-protest function or value of marijuana use for symbolizing disidentification with conventional society and the peer-acceptance function or the value of marijuana use for gaining acceptance from permissive peers. Also, the model incorporates two personality variables: general left-right sociopolitical outlook and internal sensation seeking, a propensity to seek satisfaction from fantasy, emotional change, and unusual perceptual experience (Pearson, 1970). Sociopolitical outlook should affect the symbolic-protest function because expressing disidentification with conventional society should appeal specifically to rather rebellious persons. Both personality factors should influence the peer-acceptance function because people's own attitudes and prac-

tices should predispose them to seek acceptance from similarly inclined peers (Byrne, 1971). Finally, both personality variables should affect attitude.

To test our model, we administered questionnaires to 96 male and 97 female high school seniors. The following measures were used: frequency of marijuana use over the past 6-month period; a brief, modified version of Wilson and Patterson's (1968) Conservatism Scale, which measures sociopolitical outlook; the Internal Sensation-Seeking subscale from a modified version of Pearson's (1970) Novelty Experiencing Scale (Kohn & Annis, 1975); and specially constructed measures for attitude to marijuana use, peer-acceptance function, and symbolic-protest function. All alpha or Kuder-Richardson 20 reliabilities were satisfactory, ranging from .76 to .91. In addition, the attitude and function measures in which all items referred to marijuana were shown to be factorially independent.

Path analysis, a multiple-regression procedure for testing the implications of causal models (Kim & Kohout, 1975), was applied to the data. The resultant path diagram appears in Figure 1.

The path diagram strongly resembles that implied by our model. The sole discrepancy is the absence of a significant path coefficient between internal sensation seeking and the peer-acceptance function. Possibly, internal sensation seeking is not publicly visible enough a trait to affect interpersonal attraction except in very close relationships. The residuals, symbolized by e_1 in Figure 1, express the square root of the unexplained variance for each dependent variable.

The authors gratefully acknowledge the assistance of Carol Watson and Judith Groeneveld in the various stages of data collection and analysis. They also appreciate the critical comments of Norman Endler, Michael Goodstadt, Igor Kusyszyn, and Reginald Smart on an earlier draft of this article and the advice of Michael Ornstein on path analysis.

Requests for reprints and for an extended report of this study should be sent to Paul M. Kohn, Department of Psychology, York University, Downsview, Ontario, Canada M3J 1P3.

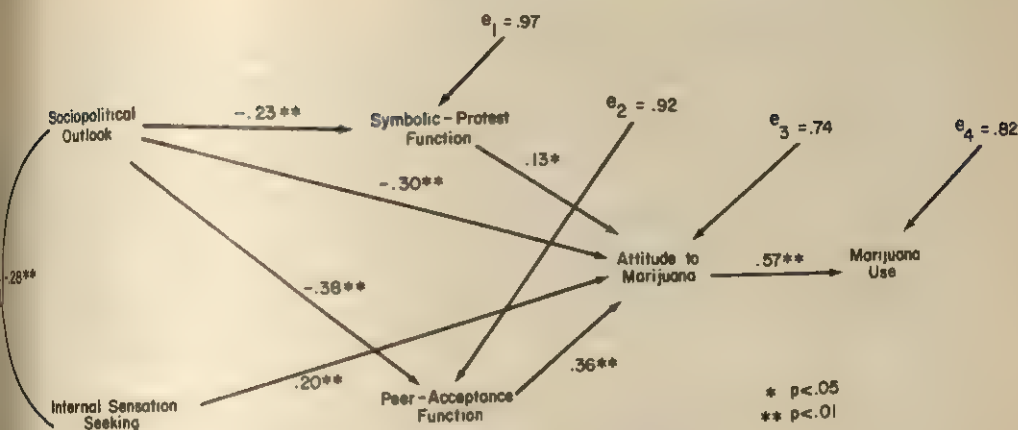


Figure 1. Path analysis for marijuana use.

Thus, the model explains about 33% of the variance in marijuana use and about 45% of the variance in attitude. The overidentification test statistic (Land, 1973) proved nonsignificant, $L(6) = 4.00$, $.70 > p > .50$. This means that our model fits the data well. Further evidence of success comes from comparisons of the observed correlations with those predicted by the model through summing direct causal, indirect causal, and noncausal contributions (Kim & Kohout, 1975). Predictive errors were generally modest and in all cases nonsignificant. Finally, the results of separate path analyses conducted for each sex separately were highly similar.

In general and with the one exception noted, the results support our a priori model. The variance accounted for in attitude and behavior, about 45% and 33%, respectively, is high enough to support the importance of the predictors selected but low enough to suggest extending the model. For example, the present model overlooks the perceived dysfunction of the various health-related and social-legal risks ascribed to marijuana use and the related personality variable of risk-taking propensity. Also, longitudinal research permitting stronger causal inferences seems desirable. Work along both these lines is proceeding.

References

Byrne, D. *The attraction paradigm*. New York: Academic Press, 1971.

Fishbein, M. Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: Wiley, 1967 (a).

Fishbein, M. A behavior theory approach to the relations between beliefs about an object and the attitude toward the object. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: Wiley, 1967 (b).

Kim, J., & Kohout, F. J. Special topics in general linear models. In N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, & D. H. Bent (Eds.), *SPSS: Statistical package for the social sciences* (2nd. ed.). New York: McGraw-Hill, 1975.

Kohn, P. M., & Annis, H. M. Validity data on a modified version of Pearson's Novelty Experiencing Scale. *Canadian Journal of Behavioural Science*, 1975, 7, 274-278.

Land, K. C. Identification, parameter estimation, and hypothesis testing in recursive sociological models. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences*. New York: Seminar Press, 1973.

Pearson, P. H. Relationships between global and specified measures of novelty seeking. *Journal of Consulting and Clinical Psychology*, 1970, 34, 199-204.

Wilson, H. D., & Patterson, J. R. A new measure of conservatism. *British Journal of Social and Clinical Psychology*, 1968, 7, 264-269.

Received May 16, 1977 ■

Weight Loss and Behavior Change 1 Year After Behavioral Treatment for Obesity

Robert W. Jeffery, Michael Vender, and Rena R. Wing
Stanford Heart Disease Prevention Program, Stanford University Medical Center

The study presents a 1-year follow-up of the first 108 clients to complete a behavioral weight reduction program at Stanford University's Eating Disorders Clinic. On the average, clients maintained their in-treatment weight loss over the follow-up period, but there was marked variability and a low correlation between in-treatment and posttreatment performance. Clients reported significant changes in their eating behavior after treatment, but these changes were only weakly related to weight changes.

Although behavioral treatments for obesity appear to produce favorable short-term results, results of follow-up studies have been weak (Stuart, 1975). Maintenance has typically been assessed over short time intervals (e.g., Hagen, 1974), and many studies suffer from small sample size or high attrition (e.g., Hall, 1973; Mahoney & Mahoney, 1976). In addition, although weight losses in behavioral programs are thought to be mediated by changes in eating behavior, there has been little research on the maintenance of new eating patterns.

The present article reports a 1-year follow-up of clients treated with behavioral techniques at the Stanford Eating Disorders Clinic (EDC), where maintenance of weight loss and behavior change are evaluated and related to the use of various weight control strategies.

Follow-up data for 88 of the first 108 clients to complete the weight control program at the EDC were derived from telephone interviews. Pretreatment weights averaged 257 pounds (116.8 kg) for males ($n = 16$) and 209 pounds (95 kg) for females ($n = 72$), with a range from 141 (64.1) to 353 pounds (160.5 kg). Time from completion of treatment varied from 12 to 18 months.

Treatment was conducted in small groups led by two experienced therapists, who met weekly for 1½ hours for 20 weeks. Lessons, presented in seminar format, covered the following topics

(Ferguson, 1975): self-monitoring, stimulus control, slowing the rate of eating, social support, exercise, diet, preplanning, and individual problem solving. Clients maintained daily records of their eating and exercise behaviors, which were checked weekly. All expenses were prepaid (\$250), and a 10% refund was given contingent on attendance and completion of homework assignments.

A subset of 25 clients completed a self-report questionnaire at the time of follow-up. In terms of demographic characteristics and weight loss, they were representative of the larger sample. They rated their eating behavior on 5-point scales at three points in time: before treatment, immediately after treatment, and currently. Twelve questions referred to behaviors taught in therapy (e.g., restricting the location of eating), 5 to thoughts about food (e.g., feeling guilty about eating), and 8 to specific eating problems (e.g., binge eating). Clients also rated the helplessness of 21 techniques presented in treatment and indicated whether each was used immediately after treatment and/or at present.

Clients lost an average of 12.8 pounds (5.8 kg) during the program and an additional .7 pounds (.3 kg) during follow-up. Ninety percent lost weight during treatment, and 43% lost additional weight during follow-up. However, weight losses by the end of follow-up were extremely variable, ranging from 80 pounds (36.4 kg) lost to 40 pounds (18.2 kg) gained. Both the percent of clients losing 20 pounds (9.1 kg) or more and the percent gaining weight more than doubled between the end of treatment and follow-up. Weight losses during follow-up were unrelated to weight losses during treatment ($r = .002$). Subject characteristics, such as sex and age of onset of obesity, were also unrelated to long-term success.

The present research was supported in part by National Institute of Mental Health Research Grant NH 26482-01 to Albert J. Stunkard.

Requests for reprints and for an extended report of this study should be sent to Robert W. Jeffery, Stanford Heart Disease Prevention Program, 730 Welch Road, Palo Alto, California 94304.

Clients reported substantial improvements in eating behavior during treatment. Twenty of the 25 questionnaire items showed significant positive changes. Although there was some deterioration during follow-up, improvements were still clearly evident in 19 items 1 year later.

Clients reported using an average of 9.4 of the 21 techniques immediately after the program, but only 6.6 by the time of follow-up, $t(22) = 3.16$, $p < .005$. The 21 behavioral techniques were grouped according to the eight topics presented in the program. In comparative analyses social support techniques were used significantly less than other techniques both immediately after treatment and at follow-up, $ts(22)$ ranged from 2.14 to 4.19, all $ps < .05$. Self-monitoring was rated significantly more helpful than any other technique, $ts(22)$ ranged from 3.28 to 5.92, all $ps < .01$. Only self-monitoring and preplanning did not decline in use significantly after treatment.

The relationship between behavioral variables and weight loss was weaker than expected. Weight loss was significantly correlated with only 5 of the 20 individual behaviors that changed during treatment and with 4 of 19 overall. Changes in thoughts about food were the best predictors of success.

Weight loss was modestly related to the reported use of behavioral techniques. Weight loss was significantly correlated with the use of self-monitoring ($r = .61$, $p < .01$) and situational restriction ($r = .47$, $p < .05$) immediately after treatment. In addition, those clients who used 11 or more techniques immediately after treatment lost more weight overall than those using fewer, $t(22) = 2.09$, $p < .05$. There were no significant relationships between weight loss and technique use at follow-up.

Weight loss following treatment in the EDC was similar to that reported in other investigations using self-control. The inability of clients

to lose additional weight on their own after behavioral treatments underscores the need for further study of maintenance processes. In the present program only 10% of the clients were within 20 pounds of their ideal weight a year after treatment and fully a third had enrolled in other weight loss programs. Almost all recommended continuing treatment contracts.

The failure to find strong relationships between self-reported behavior change and weight change, despite significant changes in eating habits, can be interpreted in several ways. Positive relationships may have been obscured by errors and biases in retrospective self-reports. Changes in eating behaviors may not be linearly related to changes in weight. Or, some clients may engage in idiosyncratic behaviors, such as traditional diets, that either enhance or retard their progress. Further research is needed to systematically assess the effects of specific changes in behavior on food consumption and weight.

References

- Ferguson, J. M. *Learning to eat*. Palo Alto, Calif.: Bull, 1975.
- Hagen, R. L. Group therapy vs. bibliotherapy in weight reduction. *Behavior Therapy*, 1974, 5, 222-234.
- Hall, S. M. Behavioral treatment of obesity: A two-year follow-up. *Behaviour Research and Therapy*, 1973, 11, 647-648.
- Mahoney, M. J., & Mahoney, K. Treatment of obesity: A clinical exploration. In B. J. Williams, S. Martin, & J. P. Foreyt (Eds.), *Obesity: Behavioral approaches to dietary management*. New York: Brunner/Mazel, 1976.
- Stuart, R. B. Behavioral control of overeating: A status report. In G. A. Bray (Ed.), *Obesity in perspective* (Report No. NIH 75-708). Washington, D.C.: U.S. Department of Health, Education and Welfare, 1975.

Received June 8, 1977 ■

Taylor's MMPI Correction Factor for Spinal Cord Injury: Empirical Endorsement

Philip C. Kendall
University of Minnesota

Jack Edinger
Virginia Commonwealth University

Carole Eberly
McGuire Veterans Administration Hospital
Richmond, Virginia

The Minnesota Multiphasic Personality Inventories (MMPIs) of 30 spinal-cord-injured (SCI) subjects, 30 hospitalized non-spinal-cord-injured controls (HC), and 30 hospital employee controls (EC) were studied here. MMPI items that significantly differentiated SCI subjects from controls were separately factor analyzed. The major factor resulting from the SCI-EC analysis was characterized by items of physical description, as was a similar SCI-HC factor. The SCI mean scale scores were then corrected for spinal cord injury by replacing the SCI response proportion for the factor-analytically identified physically descriptive items with the EC response proportion. This correction procedure was also done using Taylor's correction items. Both correction procedures resulted in significantly less pathological SCI scale means (1, 2, 3, 4, and 8). The present results support the notion that spinal cord injury can moderate MMPI item endorsement. Recommendations are discussed.

Taylor (1970) investigated the effects of spinal cord injury (SCI) as a moderator variable that affects the validity of the Minnesota Multiphasic Personality Inventory (MMPI) for persons with such injuries. Taylor contended that the physical disability affects the subject's test input by moderating the likelihood of a response endorsement and may substantially affect output and thus reduce the instrument's validity generalization. The extent to which the disability affects the MMPI is related to the number of items that the disabled patient responds to on a physically descriptive basis rather than a psychologically descriptive one.

Taylor's (1970) evaluation of SCI as a moderator variable compared the MMPIs of SCI patients with normal nonhospitalized cases and used the judgment of physical medicine special-

ists to identify items that appeared to be descriptive of the physical condition associated with spinal cord injury. Taylor then scored SCI MMPIs with a correction for the somatically relevant items (the physical descriptors). Removing the patient's response proportion for the somatically relevant items and substituting the nonpatient response proportion produced dramatic changes in profile elevation, configuration, and consequent interpretation.

Whereas Taylor (1970) used nine "specialists" to select the items that differentiated the SCI group from the nonhospitalized group, the present study intended to isolate those items that contribute to a physical rather than a psychological description by means of factor analysis. In addition, since one source of somatic concern in patients appears to be hospitalization, with its emphasis on symptoms and attention for physical difficulty, the present study includes a control for hospitalization composed of hospitalized non-SCI patients as well as nonhospitalized employee controls. Moreover, while Taylor's control group consisted of college students (average of 2 years more educated than the SCI group), the present nonhospitalized control group consisted of ward nursing aides with an educational level similar to the SCI subjects.

Three groups of male subjects, SCI ($n = 30$), hospitalized non-spinal-cord-injured subjects (HC; $n = 30$), and hospital employees (EC; $n = 30$),

Portions of this project were conducted while the first author was at Virginia Commonwealth University and the Palo Alto Veterans Administration Hospital. The authors wish to thank John Hugo for his assistance in the early conceptualization of the study and to Jeff Schulman for his assistance in data collection.

Requests for reprints and for an extended report of this study should be sent to Philip C. Kendall, Department of Psychology, Elliott Hall, 75 East River Road, University of Minnesota, Minneapolis, Minnesota 55455.

= 30) participated in the present study. The mean ages of the groups were, in the same order, 37.5, 43.4, and 37.3 years. Similarly, the mean educational levels of the groups were 11.8, 10.5, and 12.0 years. The overall mean age was 39.4 ($SD = 4.86$), and the overall educational level was 11.4 ($SD = 4.67$). Though the present groups were comparable, the present SCI group was 16.1 years older and 1.9 years less educated than Taylor's (1970) SCI group. Subjects in the present study individually completed the group form of the MMPI.

Chi-square analyses of the true-false response cells were conducted for each item in two comparisons (a) SCI-HC and (b) SCI-EC, resulting in 56 and 65 items, respectively. Similarly, Taylor reported 60 items that differentiated SCI from control groups.

Separate factor analyses were then conducted. The major factor resulting from the analysis of the items differentiating the SCI and EC groups was composed of 10 items (accounting for 26.1% of the variance) that gave a clear description of physical condition. Seven of these 10 items were among those identified by Taylor's specialists. A similar factor resulted from the analysis of the SCI-HC items, although it emerged second to a stomach-complaint factor and contained fewer items.

The SCI mean scale scores were then corrected using the EC response proportion for items factor-analytically identified and using Taylor's items. Both the factor-item-corrected and Taylor-item-corrected SCI profiles were significantly less pathological than the uncorrected SCI profiles on five scales (1, 2, 3, 4, and 8), but they remained significantly elevated when compared to the EC group profiles.

In general, the emergence of a physical description factor as the major component of

variance that distinguishes nonhospitalized patients from SCI subjects, and to a lesser extent the similar factor from the SCI-HC items, supported the validity of both (a) the items selected in the Taylor study and (b) the nature of physical condition as a moderator variable. In the present study, the factor-identified correction items (i.e., 273, 330, 9, 192, 63, 310, 51, 179, 20, 62) and Taylor's items resulted in significant reductions of the MMPI scale means (see above) of SCI patients. Even though both procedures corrected for SCI physical description, the profiles were still significantly different from the EC group.

The present study, by using a more representative group of SCI patients, more appropriate controls, and a more objective item-identification procedure, supports Taylor's contention that SCI can moderate MMPI item endorsements. To remove the unwanted effect of physical description, it is recommended that the MMPI answer sheet be scored twice, once scoring the physical description items and once deleting them. As noted by Taylor (1970), "this method specifies the minimum and maximum limits for the patient on each of the affected scales" (p. 187). Moreover, it is recommended that the empirically derived (factor-analytically identified) items reported herein constitute the correction factor.

Reference

- Taylor, G. P. Moderator-variable effects on personality-test-item endorsements of physically disabled patients. *Journal of Consulting and Clinical Psychology*, 1970, 35, 183-188.

Received June 16, 1977 ■

Meanings that Professionals Attach to Labels for Children

Charles F. Carroll
Yale University

N. Dickon Reppucci
University of Virginia

This study sought to identify relative meanings among professionals of three clinical labels for children—mentally retarded, emotionally disturbed, and juvenile delinquent—as well as an average or unlabeled choice. Forty regular classroom teachers and 32 mental health workers responded on two questionnaires designed to measure their reactions to these labels with nine questions concerned with expectations for the child's success in school and work, implications for treatment strategies, and motivation to work with the child. The labels conveyed clearly different relative meanings, and the two professional groups differed in a consistent fashion.

Recent discussion and research regarding the labeling of children (e.g., Hobbs, 1975) has pointed to undesirable consequences of labels, for example, that labels affect professionals and trainees negatively (Foster, Ysseldyke, & Reese, 1975). Nevertheless, labels continue in widespread use, since they facilitate both accountability for governmental funding and determination of eligibility for services. In addition, professionals in human service and education need shorthand methods of communicating with each other and appear to find current classification methods meaningful and helpful. The suggested use of more general labels, such as "children with special needs," is problematic not only because many more children could be so labeled but also because more general labels do not fill the governmental and professional needs just mentioned. Therefore, current labels are likely to persist, and research that focuses on questions such as the following becomes critical. What are the relative meanings attached to various labels? Is one label more positive than another? Does a label have different meanings for different professional groups?

The purposes of the present investigation were twofold: (a) to examine the relative meanings among professionals of nonlabeled or average children and three common labels for children—mentally retarded (MR), emotionally disturbed (ED), and juvenile delinquent (JD)—and (b) to determine what, if any, differences there are between teachers and mental health workers in

response to these labels. Meanings investigated included the professional's expectations for the child's success in school and work, implications for treatment and placement strategies, and motivation to work with the child. Teachers and mental health workers were studied because (a) both groups come into contact with children having a range of labels; (b) mental health workers frequently consult with teachers; and (c) children labeled by one group usually retain that label as they contact the other group.

Forty regular classroom (Grades 6-9) teachers from three middle schools and 32 mental health workers, including psychiatrists, psychologists, social workers and counselors, from four child guidance clinics and one mental health clinic participated in the study. The schools and clinics serve Connecticut areas of varied size, family income, and racial composition. Each subject completed a case study questionnaire and a labels only questionnaire. Subjects were first presented with the case study questionnaire, a *standard* six-sentence case study of a 13-year-old boy either untitled or titled with one of the three clinical labels, MR, JD, or ED. The same label, or no label, was repeated in the first sentence. Subjects answered on a 7-point scale (extremely negative to extremely positive) nine questions about the boy that reflected the three attitude areas mentioned above.

Ten teachers and eight mental health workers were randomly assigned to each of the four case study questionnaire conditions. The labels only questionnaire consisted of a repeated measures design; each subject was asked to answer the same nine questions in regard to a 13-year-old boy for each of the three clinical labels alone and for "average, without any of these prob-

Requests for reprints should be sent to N. Dickon Reppucci, Department of Psychology, Gilmer Hall, University of Virginia, Charlottesville, Virginia 22901.

lems." The first author administered both questionnaires in the subject's own work setting.

Where a significant F statistic was obtained from the least squares solution to analysis of variance ($\alpha = .05$), multiple comparisons were conducted by t test ($\alpha = .05$) or Newman-Keuls test ($\alpha = .05$ for the case study questionnaire; $\alpha = .01$ for the labels only questionnaire). On the case study questionnaire, teachers, in contrast to mental health workers, rated the children across all label conditions as less likely to finish high school, $F(1, 64) = 4.05$, $p < .05$, and less motivated to learn in school, $F(1, 64) = 4.56$, $p < .05$, and they rated themselves lower on both knowledgeable for working with the boy, $F(1, 64) = 5.40$, $p < .05$, and willingness to do so, $F(1, 64) = 6.96$, $p < .05$. Similar results were obtained on the labels only questionnaire. Teachers also rated themselves as less knowledgeable and less willing to work with all clinically labeled children as compared with the average, whereas the mental health workers rated themselves as less knowledgeable and willing only with the MR, being uniformly high for JD, ED, and average; knowledgeableness, $F(3, 210) = 13.29$, $p < .001$; willingness, $F(3, 210) = 14.24$, $p < .001$. Similar results were found regarding expectations for success at future skilled employment, $F(3, 210) = 4.06$, $p < .01$. On both questionnaires mental health workers were less approving of primarily special class placement than were teachers. For the case study questionnaire, $F(1, 64) = 3.59$, $p < .06$; for the labels only questionnaire, $F(1, 70) = 5.22$, $p < .05$. All differences for professional groups were statistically controlled for sex and age bias. In summary, labels had more negative effects for teachers than for mental health workers in the areas of professional motivation and expectations for the child's success. For mental health workers, however, MR had more negative effects than the other labels.

There were differences among the labels that were the same for the two professional groups. On the case study questionnaire, subjects reported themselves as less knowledgeable when the label MR was involved than when no label was indicated, $F(3, 64) = 2.87$, $p < .05$. The labels only questionnaire results ($df = 3, 210$, $p < .001$) indicated that subjects tended to view delinquents as less likely to finish high school ($F = 92.08$) and as less motivated to learn in school ($F = 83.35$). They also rated the JD and ED as most likely to have serious need for all

three treatments: help for his family ($F = 112.95$), someone to talk about his problems with ($F = 38.30$), and professional help to work out difficulties he has getting along with other people ($F = 52.58$). The MR were rated in the middle, and the average, lowest. There was one contradictory result in this area on the case study questionnaire: Subjects viewed the families of the MR as more in need of professional help than the families of the JD or unlabeled boys, $F(3, 64) = 3.19$, $p < .05$. Finally, on the question, "How approving are you of primarily special class placement for John?" on the case study questionnaire, the MR received more approval than the JD or unlabeled and the ED, though high, were not statistically differentiated from the other labels, $F(3, 64) = 3.64$, $p < .05$. On the labels only questionnaire, subjects rated the MR and ED highest, JD next, and average lowest, $F(3, 210) = 84.35$, $p < .001$. To summarize, professionals expressed less motivation to work with the MR, who were perceived as generally less in need of clinical services but likely candidates for special class placement. JDs were rated as low on expectations for success in school, usually as highly in need of clinical services but as inappropriate for special class placement. The ED were uniformly rated highly in need of clinical services and highly appropriate for special class placement.

Although sex differences were sparse, there were five significant ($df = 3, 210$, $p < .05$) Sex \times Label interactions on the labels only questionnaire. In essence, women were more positive than men regarding treatment for the clinically labeled children, especially the ED and JD (family help, $F = 2.77$; for talk about problems, $F = 3.19$; help in interpersonal relations, $F = 5.16$). At the same time, the women were less willing to work with the JD ($F = 3.05$) and more pessimistic about the MR finishing school ($F = 2.67$).

In conclusion, the present investigation focuses attention on three neglected aspects of the labeling of children: (a) Professionals attach different relative meanings to clinical labels, (b) teachers and mental health workers may respond differently to labeled children, and (c) professional groups may have different relative responses to common clinical labels. The study suggests that blanket indictments of a specific label may miss various subtleties that can only be determined by detailing different relative meanings of common labels. Moreover,

understanding differences between professional groups may be even more important, as children labeled by one profession frequently keep that label when they come into contact with members of other professional groups. Therefore, it is imperative to delineate both the beneficial and the negative aspects of different labels among various professional groups.

References

- Foster, G. G., Ysseldyke, J. E., & Reese, J. H. "I wouldn't have seen it if I hadn't believed it." *Exceptional Children*, 1975, 41, 469-473.
- Hobbs, N. (Ed.) *Issues in the classification of children* (2 vols.). San Francisco: Jossey-Bass, 1975.

Received July 13, 1977 ■

Effects of Anxiety and Sex on Neuropsychological Tests

Glen D. King, H. Julia Hannay, Bruce J. Masek, and Joan W. Burns
Auburn University

The effects of anxiety and sex on neuropsychological test performance were studied. Thirty male and 30 female right-handed subjects responded to the Finger Tapping (FT), Form Board (FB), and State-Trait Anxiety tests. The original cutoff points for brain damage on the neuropsychological tests resulted in numerous false positives. Females performed significantly slower on the FT than males, and for females, trait anxiety was negatively correlated with FT performance and positively correlated with time to complete the FB. The implications of these results for testing psychiatric patients are noted.

Neuropsychological tests to detect and locate brain pathology are being used increasingly by psychologists, especially in psychiatric clinics. Tests that are commonly used in neuropsychological test batteries include the Finger Tapping Test (FT) and the Form Board Test (FB). For the FT test, the subject taps an armature connected to a counter as fast as possible for 10 sec. The FB test requires that a blindfolded subject put 10 blocks of various sizes into their correct places on a formboard, first with one hand, then the other, and then both hands together. The total time across the three administrations is computed. The subject is asked to draw all of the shapes on paper in their spatial interrelationship after removal of the board. This yields a measure of memory (shapes remembered) and localization (number of correct shapes properly located).

After repeated experience with these tests in clinical and demonstration situations, we have noted that (a) performance differed by sex, but separate norms are not provided, and (b) anxiety appeared to predict poor performance. To test these notions, a study was designed using 30 male and 30 female right-handed college student volunteers with no history of brain pathology. Subjects were administered the State-Trait Anxiety Inventory and the FT and FB tests, with the order of FT and FB counterbalanced within each sex. Analyses revealed no test order effects.

The original cutoff points for brain damage (Reitan, 1955) resulted in numerous false posi-

tives across both sexes for all of the tests. However, when the performance of our subjects was compared to more recent normative data for these tests (Kløve, 1974; Vega & Parsons, 1967), we found that the number of false positives was substantially reduced.

In consideration of sex differences on FT, females ($M = 43.1$) tapped fewer times than males ($M = 47.5$) across hands, $F(1, 56) = 15.6$, $p < .01$, with the sex difference being much larger for the right hand than the left as indicated by a Sex \times Hand interaction, $F(1, 56) = 5.61$, $p < .05$. These results are inconsistent with previous data showing women to be superior on manual dexterity tasks (Maccoby & Jacklin, 1974) and warrant further study. Other comparisons for sex were not significant. Considering anxiety, there was a negative correlation between trait anxiety (A-Trait) and FT performance for females; the higher the A-Trait, the lower the number of average taps on FT ($r = -.41$, $p < .01$).

Similarly, for women only, A-Trait was positively correlated with time to complete the FB test using the preferred hand ($r = .38$, $p < .02$) and both hands together ($r = .34$, $p < .03$). It should be remembered that the subjects were college students. Only a few had clinically elevated anxiety scores, and anxiety seemed more prevalent among women than men. These results could underestimate the effects of anxiety on neuropsychological tests, especially when the tests are used with patients from a psychiatric population or with patients who suspect the presence of a serious neurologic disorder.

These results question the validity of these neuropsychological tests. For the tests to be valid, it appears that at least the FT needs to have separate norms for each sex. In addition,

Requests for reprints should be sent to Glen D. King, Department of Psychology, Auburn University, Auburn, Alabama 36830.

high anxiety in women had a deleterious effect on their performance on the FT and FB tests. This relationship needs to be carefully considered when neuropsychological tests are administered to and interpreted for psychiatric patients and neurological patients who often have high anxiety levels. The sensitivity of other neuropsychological tests to anxiety and sex differences needs to be investigated.

References

Kløve, H. Validation studies in adult clinical neuropsychology. In R. M. Reitan & L. A. Davison

- (Eds.), *Clinical neuropsychology: Current status and applications*. New York: Wiley, 1974.
- Maccoby, E. E., & Jacklin, C. N. *The psychology of sex differences*. Stanford, Calif.: Stanford University Press, 1974.
- Reitan, R. M. Investigations of the validity of Halstead's measures of biological intelligence. *A.M.A. Archives of Neurology and Psychiatry*, 1955, 73, 28-35.
- Vega, A., & Parsons, O. A. Cross-validation of the Halstead-Reitan tests for brain damage. *Journal of Consulting Psychology*, 1967, 31, 619-625.

Received July 18, 1977 ■

Diversity Among Lower-Class Therapy Clients: A Comparison of Class IV and Class V Psychotherapy Recipients

Michael T. Nietzel, Matthew G. Hile, and Charles Y. Kondo
University of Kentucky

Demographic and treatment-relevant differences were studied among 73 lower socioeconomic class psychotherapy clients at a predominantly rural comprehensive mental health center. Data were drawn from the records of all adult (46 females; 27 males) psychotherapy cases initiated and terminated in a recent 15-month period. Results indicate several differences among Social Classes IV and V despite the literature's prevalent assumption of IV-V equivalency. In addition, sufficient heterogeneity exists among Class V clients to justify a partitioning of that class into two distinct subgroups. Class IV and Class V clients also differed on a global measure of final outcome.

A substantial literature has attempted to assess the influence of client socioeconomic status (SES) on the psychotherapeutic process. One prevailing opinion shared by much of this research is an assumption of homogeneity among lower-SES clients with regard to demographic and treatment-related variables. This assumption was examined in the present study.

Subjects were persons who presented themselves for outpatient treatment at a federally supported Community Mental Health Center in rural northeast Kentucky. All cases initiated and terminated within a 15-month period were examined, and clients classified as belonging to Class IV or Class V (Hollingshead, Note 1) were selected (excluding minors and persons in special programs). The final study group consisted of 73 cases (females, $n = 46$; males, $n = 27$); 95% of the sample was white.

Using Hollingshead's weightings, three distinct groups were obtained (SES IV, $M = 55.6$; SES V_A, $M = 63.8$; SES V_B, $M = 74.2$), which t tests indicated differed at the $p < .02$ level. This suggested that Class V clients might be differentiated in this manner in subsequent analyses.

Requests for reprints and for an extended report of this study should be sent to Michael T. Nietzel, Department of Psychology, University of Kentucky, Lexington, Kentucky 40506.

Seventy-three files were examined by three psychology graduate students who made committee decisions on data drawn from the records. Demographic variables were age, sex, location of residence, education, occupation, income, marital status, number of dependent children, and type of family. Treatment-related information included source of referral, diagnosis, treatment received, number of sessions and weeks in treatment, fee, person responsible for termination, and final outcome. Contrasts of interest were tested using the chi-square statistic.

Class IV and Class V (total) clients differed significantly on occupation ($p < .001$) but not on amount of education. Class V clients were less likely to be married than Class IV clients ($p < .06$), and, if married, they were less likely to have children ($p < .05$). No differences were noted between the two groups on other demographic variables. Class IV and Class V clients did differ significantly on final outcome of therapy ($p < .02$). Class IV clients were more likely to have terminated therapy prior to their therapists' recommendations, whereas clients from Class V were more likely either to have refused service initially or to have improved (therapist evaluation) at the time of mutual termination. No other treatment variables differentiated the two groups. Comparisons of Class IV clients with Class V_A and V_B persons evidenced the same trends and differences. In addition, Class V_B received the diagnoses of schizophrenia, personality disorder, or transient situational disturbance more frequently than Class IV clients ($p < .05$).

be defined by a locus of control score of 6 or below, the more externally controlled alcoholic by a score of 7 or above. Thus, some degree of standardization is permitted, and researchers can test the means of their samples with the mean of the large normative group. The use of this base also permits ready comparison among other subpopulations of alcoholics. It is only when the most meaningful subgroups of alcoholics are so normed that useful comparisons and needed generalizations will emerge.

References

- Distefano, M., Pryer, M., & Garrison, J. Internal-external control among alcoholics. *Journal of Clinical Psychology*, 1972, 28, 36-38.
- Donovan, D., & O'Leary, M. Comparison of perceived and experienced control among alcoholics and nonalcoholics. *Journal of Abnormal Psychology*, 1975, 84, 726-728.
- Goss, A., & Morosko, T. Relation between a dimension of internal-external control and the MMPI with an alcoholic population. *Journal of Consulting and Clinical Psychology*, 1970, 34, 189-192.
- Gozall, J., & Sloan, J. Control orientation as a personality dimension among alcoholics. *Quarterly Journal of Studies on Alcohol*, 1971, 32, 159-161.
- Greenberg, E., Obitz, F., & Kaye, B. Relationships among control orientation, the FIRO-B, and the Ward Atmosphere Scale in male hospitalized alcoholics. *Journal of Studies on Alcohol*, in press.
- Gross, W., & Nerviano, V. Note on the control orientation of alcoholics. *Psychological Reports*, 1972, 31, 406.
- Obitz, F., & Oziel, L. Changes in general and specific perceived locus of control in alcoholics as a function of treatment exposure. *International Journal of the Addictions*, in press.
- O'Leary, M., Donovan, D., & Hague, W. Relationship between locus of control and defensive style among alcoholics. *Journal of Clinical Psychology*, 1975, 31, 362-363.
- Oziel, L., & Obitz, F. Control orientation in alcoholics related to extent of treatment. *Journal of Studies on Alcohol*, 1975, 36, 158-161.
- Oziel, L., Obitz, F., & Keyson, M. General and specific perceived locus of control in alcoholics. *Psychological Reports*, 1972, 30, 957-958.

Received July 20, 1977 ■

Distinguishing Learning-Disabled and Emotionally Disturbed Children on the WISC-R

Raymond S. Dean
Arizona State University

This study concerned the isolation of distinctive Wechsler Intelligence Scale for Children-Revised (WISC-R) subtest patterns that would differentiate the performance of emotionally disturbed and learning-disabled children. A stepwise discriminant analysis was used to evaluate the subtest scores of 60 learning-disabled and a matched sample of 60 emotionally disturbed children. Four subtests of the WISC-R differentiated significantly between diagnostic categories. Learning-disabled children performed predictively poorer on the Block Design, Picture Arrangement, and Object Assembly subtests and higher on Vocabulary than their emotionally disturbed counterparts. These results are interpreted as a deficit in perceptual organization for children with specific learning problems.

A problem frequently encountered by the clinician is the differentiation of emotionally disturbed and learning-disabled children. With the publication of the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) came many attempts to aid such diagnosis with subtest patterns. Despite this long-standing interest, studies dealing with diagnostic implications of WISC scores have been inconclusive.

A close scrutiny of past research dealing with characteristic WISC subtest patterns reveals certain methodological shortcomings. Of concern in the majority of studies is the univariate treatment of subtests that are multiple variates. As multiple measurements of the same individuals, subtests are interdependent and should not be split off and considered univariately. But rather, diagnostic subtest patterns would be more heuristically assessed using a multivariate procedure that considers subtests in combination. Another noteworthy concern in past research is the possible contaminating effects of inconsistently defined criteria in the selection of subjects (Dean, 1977).

The present analysis concerned the isolation of a distinct subtest pattern on the Wechsler Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974) that would discriminate between learning-disabled and emotionally disturbed children. In distinguishing between

groups, a stepwise discriminant analysis was used as a base for evaluating the predictive power of subtests. Such an analysis is used to statistically discriminate between two or more groups on the basis of a number of variables. In the present situation, groups were defined in terms of their diagnostic classification. Subtest data and IQ composites were used as the discriminating variables on which the present groups were expected to differ.

Method

The subjects were 120 Caucasian children from the Phoenix, Arizona, area who had been diagnosed by two experienced psychologists as either learning disabled or emotionally handicapped. The emotionally disturbed group, composed of 48 males and 12 females, displayed specific conduct disorders: aggressiveness, temper tantrums, anxiety states, and so forth. Moreover, the behavioral criteria proposed by Quay (1972) were used for selection. None of the children in this group was reported to have specific or generalized problems in learning. A comparable group of subjects diagnosed as learning disabled comprised the second group. The twofold basis for selection of this group was (a) the criterion recommended by Chalfant and Scheffelin (1969) for the identification of children with specific learning disabilities and (b) their goodness of match with children in the emotionally disturbed group. This match was performed with respect to sex, age (± 2 months), present grade placement, and socioeconomic status as determined by the occu-

Requests for reprints and for an extended report of this study should be sent to Raymond S. Dean, 322 Payne Hall, Arizona State University, Tempe, Arizona 85281.

pation of the family's major wage earner. Children in both groups ranged in chronological age from 6.4 to 13.6 years, with a mean of 10.06 ($SD = 3.17$) for the learning-disabled group and 10.13 ($SD = 3.21$) for the emotionally disturbed children.

Subjects were individually administered the regular subtests of the WISC-R during a 6-month period, which were scored in the usual fashion. Subtest scale scores, along with the WISC-R Verbal, Performance, and Full Scale IQ scores, were calculated for each child.

Results and Discussion

The discrepancy between the Verbal IQ and Performance IQ means for each diagnostic group was evaluated using a t comparison for dependent samples. This analysis showed the emotionally handicapped children to have a significantly higher Performance mean than that obtained on the Verbal scale, $t(119) = 16.88$, $p < .01$. The differences between Verbal and Performance IQs were not significant for learning-problem children. Hence, this finding supports the contention of Performance IQ superiority for acting-out children but not for those with learning problems, replicating the findings of Dean (1977). A 2 (sex) $\times 2$ (diagnostic group) multivariate analysis of variance was computed to test for equality of subtest and IQ composite means. The analysis revealed a significant multivariate F ratio for groups ($p < .01$). The multivariate F for sex and the Sex \times Group interaction was not significant. The significant diagnostic group multivariate F yielded significant univariate comparisons for the Block Design, $F(1, 118) = 8.43$, $p < .01$, and Object Assembly, $F(1, 118) = 5.31$, $p < .05$, subtests, with both favoring children in the acting-out group.

A stepwise discriminant analysis using the WISC-R subtests and IQ composites as discriminators was computed. The selection of variables included was based on the significance of the function as measured by overall multivariate F ratio with the addition of each variable. Using such a procedure results in the choice of an optimal set of discriminating variables.

Of the 14 possible variables, the Block Design, $F(1, 118) = 8.43$, $p < .01$, Vocabulary, $F(1, 117) = 9.45$, $p < .01$, Picture Arrangement, $F(1, 116) = 6.96$, $p < .01$, and Object Assembly, $F(1, 115) = 6.08$, $p < .01$, subtests added significantly to the amount of centroid separation between groups. The Wilk's lambda chi-square transformation showed this derived function of

variables to be highly reliable, $\chi^2(4) = 22.11$, $p < .001$.

The present results provide an interesting and somewhat unexpected view of children with learning difficulties when compared with children whose problems involve inappropriate behavior. In general, children diagnosed as learning disabled scored predictively lower on the Block Design, Picture Arrangement, and Object Assembly subtests and higher on Vocabulary than emotionally disturbed children. With the exception of Vocabulary, these subtests correspond to a WISC-R factor that Kaufman (1975) has interpreted as Perceptual Organization, and they involve visually guided motor activity. Prior evidence indicates that these subtests also require the capacity to overcome the embedding of an item from its surrounding context (Goodenough & Karp, 1961). In the main, then, children who were learning disabled appeared more field dependent in their performance on the WISC-R. This finding suggests a disturbance on the part of learning-problem children in perceptual integration, whereas children with behavior problems displayed more of a verbal deficit. Generally, learning-disabled children performed more adequately on tasks within the performance area that required verbal skills than those subtests that required nonverbal visual constructive abilities.

References

- Chalfant, J. C., & Scheffelin, M. A. (Eds.) *Central processing dysfunctions in children* (NINDS Monograph No. 9). Washington, D.C.: U.S. Government Printing Office, 1969.
- Dean, R. S. Patterns of emotional disturbance on the WISC-R. *Journal of Clinical Psychology*, 1977, 33, 486-490.
- Goodenough, D. R., & Karp, S. A. Field dependence and intellectual functioning. *Journal of Abnormal and Social Psychology*, 1961, 63, 241-246.
- Kaufman, A. S. Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology*, 1975, 43, 135-147.
- Quay, H. C. Patterns of aggression, withdrawal, and immaturity. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of a childhood*. New York: Wiley, 1972.
- Wechsler, D. *Wechsler Intelligence Scale for Children*. New York: Psychological Corporation, 1949.
- Wechsler, D. *Wechsler Intelligence Scale for Children-Revised*. New York: Psychological Corporation, 1974.

Application of Electromyographic Biofeedback to the Relaxation Training of Schizophrenic, Neurotic, and Tension Headache Patients

Frank X. Acosta and Joe Yamamoto
Department of Psychiatry and the Behavioral Sciences
University of Southern California School of Medicine

Stuart A. Wilcox
California State University, Los Angeles

This study examined the effects of electromyographic (EMG) biofeedback on tension reduction by schizophrenic, neurotic, and tension headache patients. Fourteen patients participated voluntarily in at least 10 weekly EMG biofeedback sessions at a public outpatient clinic. All had complained of chronic tension. Patients showed significant decreases in their muscle tension levels with successive biofeedback training sessions. No significant differences were found between the schizophrenic, neurotic, and tension headache groups. A further contribution was the finding that patients with diverse socioeconomic and educational levels benefited similarly from EMG biofeedback training.

Electromyographic (EMG) biofeedback of frontalis muscles has recently been reported as effective for nonpatient volunteers in treating problems of chronic tension and tension headaches (Budzynski, Stoyva, & Adler, 1970; Raskin, Johnson, & Rondestvedt, 1973). To date, no study has compared the effectiveness of EMG biofeedback on the frontalis muscles with different diagnostic groups of patients in a public psychiatric outpatient clinic. The present study was done to determine if outpatients diagnosed as schizophrenic, neurotic, or having tension headaches with serious and chronic tension would respond to EMG biofeedback.

The subjects were 14 outpatients whose clinical diagnoses were schizophrenia (6); neurosis (5); and tension headache (3). Ten of the patients were women with a mean age of 38 years; 4 were men with a mean age of 40 years. All subjects participated voluntarily in this study to learn tension reduction.

This study was presented at the annual meeting of the Western Psychological Association, Seattle, Washington, April 23, 1977.

Appreciation is extended to Lupe Guzman for her help in collecting the data.

Requests for reprints and for an extended report of this study should be sent to Frank X. Acosta, Department of Psychiatry and the Behavioral Sciences, University of Southern California School of Medicine, 1934 Hospital Place, Los Angeles, California 90033.

EMG levels were recorded in microvolts (range = 2-150 μ v) by a Model BTF-401 Biofeedback Technology, Inc., instrument. Three silver/silver chloride electrodes were placed across the forehead, approximately 1 inch (3.54 cm) above the eyebrows. A modulated tone via a table-mounted speaker provided auditory feedback to the patient. The pitch of the tone reflected the level of frontalis muscle tension.

Patients were seated in a reclining chair with eyes closed. Patients were scheduled for 1 biofeedback session per week and had to complete 10 or more sessions to be included in the study. The first visit by each patient to the biofeedback laboratory consisted of (a) taking baseline EMG measures; (b) the collection of demographic data; (c) administration of a brief IQ test, the Kent Intelligence Scale; and (d) viewing of a slide-tape cassette. Using cartoon characters and photographs, the slide tape explained the concept of biofeedback, its purpose, and how the equipment worked. Fifteen minutes of EMG baseline were recorded, which consisted of simply recording frontalis muscle EMG without any feedback being given; values were generated at 30-sec intervals. The technician recorded the values and plotted them on semilog graph paper. At the end of the baseline session and on all subsequent sessions, the subject was shown the plot of his/her EMG values and was told whether it indicated high, moderate, or low muscle tension.

The second visit by each subject was the first biofeedback training session. Each session consisted of 10 minutes of prebiofeedback EMG

baseline, during which no biofeedback tone was heard; followed by 15 minutes of biofeedback, during which the subject was instructed to reduce his/her tension by reducing the pitch of the tone from the speaker; and then 5 minutes of postbiofeedback baseline, again with no tone. Subjects were instructed to try and relax with their eyes closed but were not to fall asleep.

For data analysis, each biofeedback session was divided into four within-session measures: prebiofeedback baseline (10 minutes), first half of the biofeedback (7½ minutes), second half of the biofeedback (7½ minutes), and postbiofeedback baseline (5 minutes). Mean EMG values for each of the four measures were recorded for each biofeedback session. Several sessions were averaged to form three session blocks to even out some of the session-to-session within-subject variation. Block 1 consisted of Session 1 modified by averaging 15 minutes of no-feedback baseline of the subject's first visit with the 10 minutes of prefeedback baseline of Session 1. This gave a total of 25 minutes of prefeedback baseline covering two different visits. Block 2 consisted of averaging Sessions 4, 5, and 6. Block 3 consisted of averaging Sessions 8, 9, and 10. The study consisted of a $3 \times 3 \times 4$ factorial design, with 3 levels of diagnosis, 3 levels of session block, and 4 levels of within-session measures.

The results of a three-way analysis of variance with repeated measures showed a significant decrease ($p < .05$) in EMG levels across the three session blocks. Duncan's multiple-range test indicated that Blocks 2 and 3 both differed significantly from Block 1 ($p < .05$). This indicated that the three patient groups did lower their EMG levels with successive biofeedback training sessions. Although the schizophrenic ($M = 18.0$), neurotic ($M = 16.1$), and tension headache ($M = 29.8$) groups had different overall mean EMG levels, the analysis of variance showed no significant differences between the three diagnostic groups. No significant differences were noted for the within-session measures nor for any interactions.

To see if factors other than diagnosis may have affected the biofeedback learning, subjects

were grouped according to the following factors: intelligence, education, social class, motivation, source of referral to the study, and final disposition of the patient. Analyses of variance were performed on subject EMG levels for these factors, and no significant effects were found. However, the analyses of EMG values across Session Blocks 1, 2, and 3 were consistently significant (varying from $p < .01$ to $p < .10$). This persistent decline in EMG values across the 10 sessions further supports the findings that the patients did, in fact, lower their frontalis muscle tension with successive biofeedback training sessions.

The findings on IQ and education are of particular importance when compared with Townsend, House, and Addario's (1975) biofeedback study in a clinical setting, which screened out subjects who fell in the low IQ ranges or had less than an eighth-grade education. In contrast, our study suggests that learning to reduce tension through EMG biofeedback is applicable to individuals with diverse backgrounds.

Unlike other biofeedback studies that usually report findings based on 2-5 sessions per week (Budzynski et al., 1970; Raskin et al., 1973; Townsend et al., 1975), the patients in our study typically had only one biofeedback session per week. Interestingly, this less frequent rate of biofeedback proved to be sufficient to produce a significant reduction in frontalis muscle tension.

References

- Budzynski, T., Stoyva, J., & Adler, C. Feedback-induced muscle relaxation: Application to tension headache. *Journal of Behavior Therapy and Experimental Psychiatry*, 1970, 1, 205-211.
- Raskin, M., Johnson, G., & Rondestvedt, J. W. Chronic anxiety treated by feedback-induced muscle relaxation. *Archives of General Psychiatry*, 1973, 28, 263-267.
- Townsend, R. E., House, J. F., & Addario, D. A comparison of biofeedback-mediated relaxation and group therapy in the treatment of chronic anxiety. *American Journal of Psychiatry*, 1975, 132, 598-601.

Received August 17, 1977 ■

CHARLES C THOMAS • PUBLISHER

DIAGNOSIS AND REHABILITATION IN CLINICAL NEUROPSYCHOLOGY by Charles J. Golden, *The Univ. of South Dakota, Vermillion.* '78, 304 pp., 11 il., 13 tables, \$14.75

THE BIRTH OF LANGUAGE: The Case History of a Non-verbal Child (3rd Ptg.) by Shulamith Kastein, *Columbia Univ., New York City, and Barbara Trace.* Foreword by Sidney Carter. '77, 192 pp., 26 tables, \$8.50, paper

ADULT ASSESSMENT: A Source Book of Tests and Measures of Human Behavior by Richard S. Andrulis, *The American College, Bryn Mawr, Pennsylvania.* '77, 340 pp., \$12.50, spiral (paper)

SENSORY ISOLATION AND PERSONALITY CHANGE compiled and edited by Mark Kammerman, *Labyrinth Corporation, Beverly Hills, California.* (13 Contributors) '77, 324 pp., 2 il., 39 tables, \$12.75

PSYCHIATRIC DRUGS FOR THE NON-MEDICAL MENTAL HEALTH WORKER by William Goldsmith, *Brentwood Veterans Administration Hospital, Los Angeles.* '77, 188 pp., 27 il., 7 tables, \$11.50

HOMEWORK IN COUNSELING AND PSYCHOTHERAPY: Examples of Systematic Assignments for Therapeutic Use by Mental Health Professionals (2nd Ptg.) by John L. Shelton, *Colorado State Univ., Fort Collins, and J. Mark Ackerman, Private Practice, Albany, Oregon.* '76, 308 pp., 27 il., 1 table, cloth-\$16.75, paper-\$12.95

GAMBLING AND SOCIETY: Interdisciplinary Studies on the Subject of Gambling edited by William R. Eadington, *Univ. of Nevada, Reno.* (34 Contributors) '76, 488 pp., 27 il., 40 tables, \$34.75

THE PSYCHOLOGY OF LEISURE: Research Approaches to the Study of Leisure (2nd Ptg.) by John Neulinger, *City College of the City Univ. of New York, New York City.* '76, 240 pp., 1 il., 14 tables, \$10.50

BASIC VISUAL PROCESSES AND LEARNING DISABILITY edited by Gerald Leisman, *Brooklyn College of the City Univ. of New York, Brooklyn.* (18 Contributors) '76, 456 pp. (6 3/4 x 9 3/4), 136 il., 18 tables, \$27.50

BASIC PSYCHIATRY FOR CORRECTIONS WORKERS by Henry L. Hartman, *Court Diagnostic and Treatment Center, Toledo, Ohio.* '78, 488 pp., 3 il., 1 table, \$16.75

PRESCRIPTIONS FOR CHILDREN WITH LEARNING AND ADJUSTMENT PROBLEMS (5th Ptg.) by Ralph F. Blanco, *Temple Univ., Philadelphia.* '77, 320 pp., \$10.50

LEISURE COUNSELING: An Aspect of Leisure Education edited by Arlin Epperson, *Univ. of Missouri, Columbia; Peter A. Witt, Univ. of Ottawa, Ottawa, Ontario, Canada; and Gerald Hitzhusen, Univ. of Missouri, Columbia.* (31 Contributors) '77, 392 pp., 16 il., 22 tables, \$18.00

TRAFFIC ENVIRONMENT AND THE DRIVER: Driver Behavior and Training in International Perspective by Julius Marek, *Univ. of Bergen, Bergen, Norway, and Terje Sten, Univ. of Trondheim, Trondheim, Norway.* Foreword by Karsten Krogsaeter. '77, 284 pp., 106 il., 16 tables, cloth-\$16.75, paper-\$12.95

MANAGING THE SEVERELY RETARDED: A Sampler compiled and edited by David Gibson, *Univ. of Calgary, and Roy I. Brown, The Vocational and Rehabilitation Research Institute, both of Calgary, Alberta, Canada.* (95 Contributors) '76, 500 pp., 34 il., 53 tables, \$22.75

SCHOOL CONSULTATION: Readings about Preventive Techniques for Pupil Personnel Workers edited by Joel Meyers, Roy Martin and Irwin Hyman, *all of Temple Univ., Philadelphia, Pennsylvania.* (29 Contributors) '77, 368 pp., 8 il., 20 tables, \$14.95, paper

ACUTE GRIEF AND THE FUNERAL edited by Vanderlyn R. Pine, Austin H. Kutscher, David Peretz, Robert C. Slater, Robert DeBellis, Robert J. Volk, and Daniel J. Chericco. '76, 328 pp., 1 il., 26 tables, \$16.50

BECOMING A PSYCHOTHERAPIST by Jacquelin Goldman, *Univ. of Florida, Gainesville.* '76, 140 pp., 2 il., \$9.50

THE CHANGING STATUS OF THE BLIND: From Separation to Integration by Berthold Lowenfeld, *Berkeley, California.* '75, 352 pp., 5 tables, \$23.50

Orders with remittance sent, on approval, postpaid

301-327 East Lawrence Avenue • Springfield • Illinois • 62717

Psychologists . . .
Researchers . . .
Teachers . . .
Students . . .

NOW at pre-publication prices!
(Until July 31, 1978)

Cumulative Subject and Author Indexes to Psychological Abstracts 1975-1977

*Summarizing over 77,000 articles in
three handsome, permanent volumes*

Published by The American Psychological Association

The three volumes of this vital reference tool include cumulative indexes by subject and author of all the material included in 36 issues of *Psychological Abstracts* . . . a total of over 77,000 articles. Covering the three-year period from January 1975 through December 1977, the index includes not only the material previously published in the six semi-annual indexes but also additional bibliographic data making it easier to find both original sources and abstracted material.

These permanent volumes, created to withstand years of active use, are approximately 10" x 14", printed on durable, acid-free paper and handsomely encased in Class A Library bindings.

With one volume referencing material by author name and two by subject matter, the cumulative indexes will facilitate research for PA subscribers and students, researchers and others interested in psychological literature.

PRICES Cumulative Indexes to Psychological Abstracts 1975-1977

	US	Non-US
<i>If ordered before July 31, 1978</i>		
Subject Index only (2 volumes)	\$365	\$370
Author Index only (1 volume)	\$225	\$230
Both Indexes	\$450	\$460
<i>If ordered after July 31, 1978</i>		
Subject Index only	\$435	\$445
Author Index only	\$275	\$285
Both Indexes	\$550	\$560

Prices for Previous Editions Cumulative Indexes to Psychological Abstracts 1972-1974

Subject Index only (2 volumes)	\$380	\$390
Author Index only (1 volume)	\$235	\$245
Both Indexes	\$485	\$495

Cumulative Indexes to Psychological Abstracts 1969-1971

Subject Index only (2 volumes)	\$360	\$390
Author Index only (1 volume)	\$220	\$240
Both Indexes	\$465	\$475

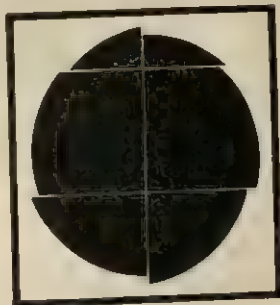
Cumulative Indexes to Psychological Abstracts for the period 1927-1968 may be obtained by writing to: G. K. Hall & Company, 70 Lincoln Street, Boston, MA 02111. Please do not write to APA for information about these earlier editions.

ORDERING INFORMATION

Please send your request for the 1975-1977 INDEX accompanied by full payment or institutional purchase order to:

American Psychological Association
Publication Sales Department
1200 17th Street, N.W.
Washington, D.C. 20036

Evaluation of Psychological Service Delivery Programs



Special Issue of Professional Psychology

Evaluation of Psychological Service Delivery Programs is the title of a special November, 1977 issue of Professional Psychology. Editor Allan Barclay and guest Co-editors Bob and Evelyn Perloff have assembled a distinguished collection of articles that brings readers up to date on the evaluation of psychological services. The issue addresses the need to justify the effectiveness of service delivery programs not only in terms of cost but more importantly in terms of human benefits.

Sample Topics and Contributors

A sample of the topics and contributors appearing in this issue includes: Theodore Blau on quality of life; Nicholas Cummings on psychotherapy; Howard Blane on alcoholism; S. B. Sells and associates on drug abuse; and NIMH's Bertram Brown, Benjamin Liptzin and James Stockdill on the federal view of mental health program evaluation.

Ordering Information

Single copies of the issue are available from the Order Department at APA for \$5 each. All orders \$25 or under must be prepaid. Please make checks payable to APA. Send your orders to:



Order Department, P.P.
American Psychological
Association
1200 17th Street, N.W.
Washington, D.C. 20036

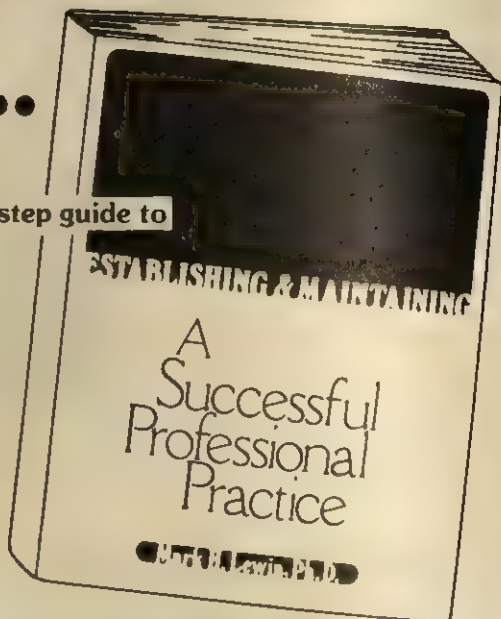
For the First Time...

an easy reading, practical, step by step guide to

A must for

- Full and Part-time private practitioners.
- Those who have ever even considered entering private practice.

**"Reduces the chance of
failure or disappointment."**



An entertaining, well organized book complete with worksheets, plans, and practical approaches to the everyday "nuts and bolts" business aspects of private practice.

MARK H. LEWIN, an ABPP Diplomate, distinguished author and editor, and founder of a thriving independent psychological service center, gives step-by-step procedures for:

- Assessing your readiness for private practice
- Preplanning the practice
- Setting up the office
- Developing new referrals
- Managing your time
- Effective use of lawyers and accountants
- Evaluating the practice as it develops
- Budgeting, fee setting, controlling costs and tax expenses

First Printing 1978 - \$25.00
10% Discount Available Now
\$22.50

Mail checks payable to:
Professional Development Institute
756 East Main Street
Rochester, N.Y. 14605

Please send me _____ copy(ies) of **Establishing and Maintaining a Successful Professional Practice** for which I enclose \$22.50 per book plus New York sales tax where applicable.

Name _____

Street Address _____

City _____ State _____ Zip _____

TAKE TIME



As an APA member,
your time is worth up to \$100 a day.

Hospital Indemnity

pays you \$40, \$60, \$80 or \$100 a day — every day that you are hospitalized. Double the amount when you are in an Intensive Care Unit.

A full year of benefits — with no waiting period. Your family can be covered too.

Your time in the hospital is worth money.
Money you can use any way you want.



Because your time is worth money.

Take time and mail this coupon today.

Mr. Harold Belcher, Administrator
APA Insurance Plans
Liberty Mutual Insurance Company
175 Berkeley Street
Boston, Massachusetts 02117

**LIBERTY
MUTUAL**

LIBERTY MUTUAL INSURANCE COMPANY • BOSTON



Yes! I would like more information about APA — sponsored
Hospital Indemnity Insurance.

Name _____

Print

Address _____

City _____

State _____

Zip _____

7/12
78

Could you use objectively obtained estimates of strengths and interrelationships of the id, ego, and superego in your diagnostic work? In your research on personality? Are you just plain curious about how these basic concepts of psychoanalytic theory might be quantified?

The I E S TEST

utilizes four independent subtests to measure important aspects of id, ego, and superego functioning. Materials are easy to handle. Tasks are disarming and are intrinsically interesting to almost anyone. Evaluation takes as much time, thought, and sophistication as for any other test giving important basic information about the personality.

Obtain information and/or order (complete set \$43.50) from

**PSYCHOLOGICAL TEST
SPECIALISTS**

Box 9229

Missoula, Montana 59807

VOLUME 46, NUMBER 3

JUNE 1978

Handwritten:
M 394

JOURNAL OF
CONSULTING
AND
CLINICAL
PSYCHOLOGY

AMERICAN PSYCHOLOGICAL ASSOCIATION
ISSN 0022-006X

Editor

Brendan A. Maher, *Harvard University*

Associate Editor

Judith P. Worell, *University of Kentucky*

Consulting Editors

Thomas M. Achenbach, *National Institute of Mental Health, Bethesda, Maryland*

Kent G. Bailey, *Virginia Commonwealth University*

Barbara B. Brown, *Beverly Hills, California*

James N. Butcher, *University of Minnesota*

Jean Chapman, *University of Wisconsin—Madison*

George M. Chartier, *Arizona State University*

Andrew L. Comrey, *University of California, Los Angeles*

W. Grant Dahlstrom, *University of North Carolina, Chapel Hill*

Anthony Davids, *Brown University*

Gerald C. Davison, *State University of New York at Stony Brook*

Alan S. DeWolfe, *Loyola University of Chicago*

Barbara Dohrenwend, *Graduate School and University Center of the City University of New York*

Richard M. Eisler, *Virginia Polytechnic Institute and State University, Blacksburg, Virginia*

Norman S. Endler, *York University, Downsview, Canada*

Sol Garfield, *Washington University in St. Louis*

Donna M. Gelfand, *University of Utah*

Goldine C. Gleier, *University of Cincinnati Medical Center*

Harrison G. Gough, *University of California, Berkeley*

Jesse G. Harris, Jr., *University of Kentucky*

Mary B. Harris, *University of New Mexico*

Kenneth Heller, *Indiana University*

Frances A. Hill, *University of Montana*

Alexandra G. Kaplan, *University of Massachusetts—Amherst*

Alan S. Kaufman, *University of Georgia*

Alan E. Kazdin, *Pennsylvania State University*

Jeffrey A. Kelly, *University of Mississippi Medical Center*

Barbara K. Keogh, *University of California, Los Angeles*

James C. Kincannon, *University of Minnesota*

Richard I. Lanyon, *Arizona State University*

Herbert M. Lefcourt, *University of Waterloo, Waterloo, Canada*

Peter M. Lewinsohn, *University of Oregon*

Michael J. Mahoney, *Pennsylvania State University*

G. Alan Marlatt, *University of Washington*

Albert Marston, *University of Southern California*

Eric J. Mash, *University of Calgary, Calgary, Canada*

Richard McFall, *University of Wisconsin—Madison*

Martha T. S. Mednick, *University of Connecticut*

Barbara G. Melamed, *University of Florida*

David I. Mostofsky, *Boston University*

Benjamin J. Murawski, *Peter Bent Brigham Hospital, Boston, Massachusetts*

Peter E. Nathan, *Rutgers—The State University*

John M. Neale, *State University of New York at Stony Brook*

K. Daniel O'Leary, *State University of New York at Stony Brook*

Oscar A. Parsons, *University of Oklahoma*

Nolan E. Penn, *University of California, San Diego*

N. Dickon Reppucci, *University of Virginia*

Jerome Resnick, *Temple University*

Leonard G. Rorer, *Miami University*

Gerald M. Rosen, *Providence Family Medical Center, Seattle, Washington*

Robert Rosenthal, *Harvard University*

William Schofield, *University Hospitals, Minneapolis, Minnesota*

Carolyn Simmons, *University of Colorado at Denver*

Bonnie Strickland, *University of Massachusetts—Amherst*

Hans H. Strupp, *Vanderbilt University*

Richard M. Suinn, *Colorado State University*

Richard S. Swit, *Duke University Medical Center*

Patricia B. Sutker, *Medical University of South Carolina*

Norman F. Watt, *University of Massachusetts—Amherst*

Robert L. Weiss, *University of Oregon*

Joan Welkowitz, *New York University*

G. Terence Wilson, *Rutgers—The State University*

Patricia Wisocki, *University of Massachusetts—Amherst*

Melvin Zax, *University of Rochester*

APA Journal Staff

Anita DeVivo, *Executive Editor*

Deborah S. Felder, *Assistant to the Executive Editor*

Kendall C. Patterson, *Production Supervisor*

Barbara R. Richman, *Production Editor*

Robert J. Hayward, *Advertising Representative*

Juanita Brodie, *Subscription Manager*

The *Journal of Consulting and Clinical Psychology* publishes original contributions on the following topics: (a) the development, validity, and use of techniques of diagnosis and treatment in disordered behavior; (b) studies of populations of clinical interest, such as hospital, prison, rehabilitation, geriatric, and similar samples; (c) cross-cultural and demographic studies of interest for the behavior disorders; (d) studies of personality and of its assessment and development where these have a clear bearing on problems of consulting and clinical psychology; or (e) case studies pertinent to the preceding topics. In order to improve the use of journal resources it has been agreed by the two Editors concerned that the *Journal of Consulting and Clinical Psychology* will no longer consider manuscripts dealing with the etiology or descriptive pathology of abnormal behavior, and the *Journal of Abnormal Psychology* will no longer consider articles dealing with the diagnosis or treatment of abnormal behavior. Articles that appear to have a significant contribution to both of these broad areas may be sent to either journal for editorial decision. Papers of a theoretical nature will also be considered within the space limitations of the journal.

Manuscripts: Submit manuscripts in triplicate to the Editor-elect, Sol L. Garfield, Department of Psychology, Washington University, St. Louis, Missouri 63130 according to instructions provided elsewhere in the journal. Authors submitting a manuscript previously considered for publication in another APA journal are invited to inform the Editor, who will then seek to obtain independent reviews, thus avoiding the possibility of repeated reviewing by the same consultant.

Copyright and Permission: All rights reserved. Written permission must be obtained from the American Psychological Association for copying or reprinting text of more than 500 words, tables, or figures. Permissions are normally granted contingent upon like permission of the author, in addition to the APA copyright notice on the first page of reproduced material, and payment of a fee of \$10 per page, table, or figure of APA material. Permission and fees are waived for authors who wish to reproduce their own material for personal use. Fees only are waived for authors who wish to use their own material commercially. Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. copyright law: (1) those post-1977 articles that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through the Copyright Clearance Center, P.O. Box 765, Schenectady, NY 12301; (2) pre-1978 articles without fee. Instructors and educational institutions are permitted to photocopy isolated articles for noncommercial classroom or library reserve use without fee. Address requests for permission to the Permissions Office, APA, 1200 Seventeenth Street, N.W., Washington, D.C. 20036.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1978: \$32 domestic, \$34 foreign, \$6 single issue. APA member rate: \$16. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Department, APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P.O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Published bimonthly (beginning in February) in one volume per year by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D.C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D.C., and at additional mailing offices.

Journal of
Consulting and Clinical Psychology

Copyright © 1978 by the American Psychological Association, Inc.

June 1978 Volume 46, Number 3

Articles

- 385 Effects of Dopamine Synthesis Inhibition on WAIS Comprehension/Edward F. Donnelly, Henry A. Nasrallah, Richard Jed Wyatt, J. Christian Gillin, and Llewellyn B. Bigelow
- 389 Psychophysiological Effects of Progressive Relaxation in Anxiety Neurotic Patients and of Progressive Relaxation and Alpha Feedback in Nonpatients/Paul M. Lehrer
- 405 Objective Measurement of Fear of Success and Fear of Failure: A Factor Analytic Approach/Susan Sadd, Michael Lenauer, Phillip Shaver, and Noel Dunivant
- 417 WISC-R Factor Structures Among Anglos, Blacks, Chicanos, and Native-American Papagos/Daniel J. Reschly
- 423 General Versus Specific Traits in the Assessment of Anxiety/Martin Mellstrom, Jr., Marvin Zuckerman, and George A. Cicala
- 432 Effect of Subliminal Stimulation of Symbiotic Fantasies on Behavior Modification Treatment of Obesity/Lloyd H. Silverman, April Martin, Roseann Ungaro, and Eric Mendelsohn
- 442 Specific and Nonspecific Factors in the Effectiveness of a Behavioral Approach to the Treatment of Marital Discord/Neil S. Jacobson
- 453 Toward the Assessment of Social Competence/Robert W. Levenson and John M. Gottman
- 463 Behavior Modification for Obesity: The Evaluation of Exercise, Contingency Management, and Program Adherence/Peter M. Stalonas, Jr., William G. Johnson, and Maryann Christ
- 470 A Longitudinal Study of the Personality Correlates of Marijuana Use/Edwin J. Kay, Arthur Lyons, William Newman, Donald Mankin, and Roger C. Loeb
- 478 The Child Behavior Profile: I. Boys Aged 6-11/Thomas M. Achenbach
- 489 Increasing the Interpersonal Problem-Solving Skills of an Alcoholic Population/James C. Intagliata
- 499 Preparation for a Stressful Medical Procedure: Effect of Amount of Stimulus Pre-exposure and Coping Style/Robert H. Shipley, James H. Butt, Bruce Horwitz, and John E. Farbry
- 508 Addict Descriptions of Therapeutic Community, Multimodality, and Methadone Maintenance Treatment Clients and Staff/Patricia B. Sutker, Albert N. Allain, Charles J. Smith, and Gary H. Cohen
- 518 Treatment of General Tension: Subjective and Physiological Effects of Progressive Relaxation/T. D. Borkovec, J. B. Grayson, and K. M. Cooper
- 529 Behavioral Contingencies and Self-mutilation in Lesch-Nyhan Disease/Lowell Anderson, Joseph Dancis, and Murray Alpert
- 537 On the Decision to be Assertive/Decky Fiedler and Lee Roy Beach
- 547 Use of Paradoxical Intention in a Behavioral Program for Sleep Onset Insomnia/L. Michael Ascher and Jay S. Efran
- 551 A Scoring Manual for Literalness in Proverb Interpretation/Chris A. Hertler, Loren J. Chapman, and Jean P. Chapman

(Continued on next page)

Case Study

- 556 Cohesive and Dispersal Behaviors: Two Classes of Concomitant Change in Psychotherapy/*Leonard M. Horowitz, Harold Sampson, Ellen Y. Siegelman, Joseph Weiss, and Shirley Goodfriend*

Brief Reports

- 565 Locus of Control, Prediction, and Performance on University Examinations/*Timothy M. Gilmor and David W. Reid*
- 567 Person \times Situation Interactions in Personality Prediction: Some Specifics of the Person Factor/*Lawrence J. Breen, Norman S. Endler, Terry J. Prociuk, and Marilyn Okada*
- 569 Abnormality of Subtest Score Differences on the WISC-R/*Richard J. Piotrowski*
- 571 Correlation of WAIS IQ in 10 Pairs of Brothers/*Joseph D. Matarazzo, Arthur N. Wiens, and Allen E. Shealy*
- 573 Self-regulatory Effects of Monitoring Sensory and Affective Dimensions of Pain/*Kenneth D. Craig, Helen Best, and J. Allan Best*
- 575 Comparison of the Similarities and Differences in the Self-concepts of Male Alcoholics and Addicts/*Jerome F. X. Carroll, M. Israel Klein, and Yoav Santo*
- 577 Dimensions, Patterns, and Personality Correlates of Drug Abuse in an Offender Population/*Terrill R. Holland*
- 579 MMPI Evaluation of 5-Year Methadone Treatment Status/*Gennaro Ottomanelli, Peter Wilson, and Richard Whyte*
- 582 Stimulus-Seeking Behavior in Three Delinquent Personality Types/*David A. Shostak and Curtis W. McIntyre*
- 583 Validity Generalization of the WISC-R Factor Structure with 10½-Year-Old Children/*David A. Shiek and John E. Miller*
- 584 Comparison of Self-administered and Examiner-administered Depression Adjective Check Lists/*Bernard Lubin, Joseph G. Marone, and Ronald G. Nathan*
- 586 Early Childhood Autism and Structural Therapy: Outcome After 3 Years/*Alan J. Ward*
- 588 Coital Position and Sex Roles: Responses to Cross-sex Behavior in Bed/*Elizabeth Rice Allgeier and Arthur F. Fogel*
- 590 Variations in a Construct: Quantitative and Qualitative Differences in Children's Locus of Control/*Philip C. Kendall, A. J. Finch, Jr., Verda L. Little, Bernard M. Chirico, and Thomas H. Ollendick*
- 593 Therapist and Group Contact as Variables in the Behavioral Treatment of Obesity/*Kelly D. Brownell, Carol L. Heckerman, and Robert J. Westlake*

Other

- 416 Garfield Appointed Editor, 1979-1984
- 431 Instructions to Authors

Effects of Dopamine Synthesis Inhibition on WAIS Comprehension

Edward F. Donnelly, Henry A. Nasrallah, Richard Jed Wyatt,
J. Christian Gillin, and Llewellyn B. Bigelow

National Institute of Mental Health, St. Elizabeths Hospital, Washington, D.C.

A previous study has shown that parkinsonian patients treated with L-dopa had decreased scores over time on the Comprehension subtest of the Wechsler Adult Intelligence Scale (WAIS), whereas scores on all other 10 subtests of the WAIS increased. It was hypothesized that if L-dopa treatment, which increases dopamine activity in the brain, is directly related to an apparent deleterious effect on the WAIS Comprehension, then a drug such as alpha-methyl-*para*-tyrosine (AMPT), which decreases dopamine activity, might have an augmenting effect on this subtest. A therapeutic trial of AMPT in a group of chronic schizophrenic patients provided an opportunity to test this hypothesis. It was found that Comprehension scores improved significantly with AMPT. Other clinical rating instruments failed to show any changes. The implications of using a psychometric instrument to assess specific, but clinically obscured, drug effects on intellectual functioning are discussed.

Over the past 25 years, treatment of the schizophrenic disorders has been dramatically altered with the widespread use of antipsychotic drugs. A large number of studies have established that chlorpromazine and other antipsychotic drugs are effective for the treatment of acute psychotic decompensation and the prevention of relapses (Engelhardt, Freedman, Rosen, Mann, & Margolis, 1964; Lasky, Klett, Caffey, Bennett, Rosenblum, & Hollister, 1962; National Institute of Mental Health, 1964).

Frequent side effects of the treatment of schizophrenic patients with antipsychotic drugs are tremors, muscle rigidity, stiff posture, shuffling gait, and drooling. Because of the similarity of these side effects to the symptoms of Parkinson's disease, they have been referred to as "parkinsonian" side effects. Further, it has been shown that these side effects in schizophrenic patients are related to the blockade of dopamine receptors

by antipsychotic drugs (Snyder, Banerjee, Yamamura, & Greenberg, 1974). Research in parkinsonism itself has revealed that some brain areas contributing to the control of body movements have lower concentrations of the neurotransmitter dopamine than normals. When parkinsonian patients were treated with L-dopa (dopamine itself cannot be used in replacement therapy because it does not cross between the blood stream and brain tissue, whereas its synthetic precursor, L-dopa, does), their symptoms often disappeared or became less severe.

Improvement of schizophrenic symptoms seems to be associated with drugs that decrease brain dopamine activity, whereas improvement in parkinsonian symptoms is associated with drugs that increase dopamine activity in the brain.

It is generally agreed that the Comprehension subtest of the Wechsler Adult Intelligence Scale (WAIS) is a measure of some form of "judgment." A recent study (Donnelly & Chase, 1973) of parkinsonian patients placed on therapeutic trials of L-dopa reported a decrease in the Comprehension scores after several months of treatment; in con-

Requests for reprints should be sent to Edward F. Donnelly, National Institute of Mental Health, St. Elizabeths Hospital, William A. White Building, Washington, D.C. 20032.

In the public domain.

trast, scores on all other 10 subtests of the WAIS showed an increase. A possible drug-induced defect in this type of judgment is consistent with exacerbation of psychosis in chronic schizophrenic patients after L-dopa treatment (Yaryura-Tobias, Wolpert, Dana, & Merlis, 1970) and with other reports of agitation, confusion, paranoia, delusions, and depression in parkinsonian patients after L-dopa treatment (Cotzias, Papavasiliou, & Gellene, 1969; Yahr, Duvoisin, Shear, Barrett, & Hoehn, 1969; Barbeau, 1972). These findings, suggesting that L-dopa is detrimental to cognitive functioning, led us to hypothesize that depleting the brain of its dopamine stores might be accompanied by an increase in the scores of the Comprehension subtest. Alpha-methyl-*para*-tyrosine (AMPT), a drug known to inhibit the rate-limiting enzymatic step in dopamine synthesis (Spector, Sjoerdma, & Udenfriend, 1965), was used to test our hypothesis.

The main objective of the present study, a part of a more extensive study of the potential clinical effectiveness of AMPT in schizophrenia (Nasrallah et al., 1977), was to determine if the WAIS Comprehension scores of a small but relatively homogeneous group of chronic schizophrenic patients, already receiving high doses of antipsychotic drugs (e.g., chlorpromazine, fluphenazine, and thioridazine), would increase with the administration of AMPT.

Method

Subjects

Nine voluntary male chronic schizophrenic patients, selected for their poor response to conventional antipsychotic drugs, who were hospitalized on a research ward at St. Elizabeths Hospital, Washington, D.C., participated in the present study. They ranged in age from 19 to 35 years, with a mean of 26.2 years. Length of hospitalization ranged from 2 to 15 years, with a mean of 8.3 years. Each patient satisfied the research diagnostic criteria (Spitzer, Endicott, & Robins, 1975) for chronic undifferentiated schizophrenia.

Procedure

The patients had been on the research ward for 4 to 13 months when consent was obtained from them and their families to participate in this study. Although the patients were not on optimal dosages

of medication at the time of the study, experience indicated that greater doses of antipsychotic drugs would not have enabled them to leave the hospital for a sustained period of time. Each of the patients had been stabilized on a constant dose of antipsychotic drugs (average equivalent dose of chlorpromazine = 800 mg/24 hr) for several months before the study, and all were kept on the same dose throughout the study. Patients entered the protocol over a 12-month period as they became available.

AMPT and placebo were administered in a double-blind nonrandom design, with periods on drug and placebo for each patient. A 4-week period of placebo administration preceded and followed a 3-week trial of AMPT. A fixed number of capsules (a total of 12 placebo, placebo plus AMPT, or AMPT capsules alone) was given daily to each patient in four divided doses. To assess each patient's clinical status, the Comprehension subtest of the WAIS was administered and scored by the same psychologist at the end of the first placebo period, the AMPT, and the last placebo period. Two other psychologists, blind to the design of the study, also scored this subtest (intrater reliability $W = .939$).

The patient's clinical status was also monitored daily by the National Institute of Mental Health (NIMH) Inpatient Behavioral Rating Scale, a 26-item nurses' rating scale characterizing a broad range of psychopathology and showing good interrater reliability (Green, Bigelow, O'Brien, Stahl, & Wyatt, 1977). Nurses completing this form were blind to the experimental design and medication status of the patients. The Brief Psychiatric Rating Scale (BPRS; Overall & Gorham, 1962) was completed by a psychiatrist who also was blind to design and medication status.

Results

WAIS Comprehension Subtest Scores

A Friedman two-way analysis of variance indicated significant differences ($\chi^2 = 11.56$, $p < .001$) between the AMPT trial and the two placebo periods. The Wilcoxon matched-pairs signed-ranks test was then used to determine the differences between the AMPT trial and the two placebo periods. Means and standard deviations from the first placebo period to the AMPT trial were 6.56 ± 2.54 , and 10.33 ± 2.16 , respectively ($T = 1$, $p < .01$). Means and standard deviations from the AMPT trial and the second placebo period were 10.33 ± 2.16 and 8.00 ± 2.26 , respectively ($T = 0$, $p < .01$). However, there was no significant difference between the first and second placebo periods ($T = 10$). Inspection of the individual scores of the nine patients in Table 1 shows that every patient got worse

during the second placebo period, and all but one improved during the AMPT trial.

Other Patient Ratings

The means of the psychosis scores of the nurses' NIMH Inpatient Behavioral Rating Scale did not show significant changes among the first placebo, AMPT, and second placebo periods (11.42 ± 3.11 , 11.93 ± 2.54 , and 11.23 ± 3.96 , respectively). Similarly, the BPRS schizophrenia scores did not show significant changes among the first placebo, AMPT, and second placebo periods (9.54 ± 3.39 , 7.56 ± 3.11 , and 8.43 ± 3.11 , respectively).

Discussion

The findings of this study support the hypothesis that depleting the brain of its dopamine stores produces an increase in WAIS Comprehension scores. Thus, judgment, as measured by the Comprehension subtest, in schizophrenic patients appears to improve in a relatively "hypodopaminergic" state. Moreover, our results also suggest that Comprehension may be a more sensitive measure of subtle drug-induced intellectual changes in schizophrenic patients than the nurses' rating scale (NIMH Inpatient Behavioral Rating Scale) or the interview-based BPRS.

The possibility of artifacts arising consequent to repeated testing in each patient at relatively short intervals must be considered. It is conceivable that the increase in scores at the time of the second testing (AMPT trial) was a function of a practice effect. Likewise, the decrement in scores observed at the third testing (second placebo trial) could be attributed to a boredom effect. Such a coincidence of influences, however, seems unlikely.

The possibility that improvement in the Comprehension scores was due to a placebo effect also must be considered. If it is due to such an effect, it would have to be operative only between the 4th and 7th week of the protocol and be sufficiently evanescent to disappear at the time of the third testing, given at the close of the 11th week of the study.

Results could not be due to a group trend or overall shifts in the ward environment, be-

Table 1
Individual Comprehension Scores (Raw) of Nine Chronic Schizophrenic Patients on AMPT and Two Placebo Periods

Patient	First placebo	AMPT	Second placebo
1	0	7	6
2	6	11	10
3	8	7	4
4	9	13	11
5	6	9	7
6	8	11	9
7	6	10	7
8	8	12	7
9	8	13	11

Note. AMPT = alpha-methyl-*para*-tyrosine.

cause the patients were run at different times, not as a cohort.

To assess whether administration of AMPT affected central dopamine in this study, we measured plasma prolactin levels. The release of this hormone, secreted by the pituitary gland, is normally inhibited by dopamine. Since serum prolactin levels did increase significantly during this study, it appears likely that AMPT did inhibit synthesis of dopamine in the brain (Nasrallah, Rogol, Wyatt, & Gillin, Note 1).

These results, together with the earlier finding that parkinsonian patients performed worse on the Comprehension subtest when treated with L-dopa than without, suggest that performance on this subtest is inversely related to central dopaminergic activity.

At the present time, assessment of drug-induced changes in psychiatric populations relies heavily on rating scales, symptom checklists, and self-reports. Our data, however, suggest that the relationship between a specific intellectual function and the specific brain neurotransmitter dopamine can be most accurately determined by quantitative assessment, that is, the patient's scored responses to the Comprehension subtest. Several studies that have investigated the effects of antipsychotic drugs on the WAIS scores of normal and schizophrenic subjects reported significantly improved scores (Abrams, 1958; Gardner, Hawkins, Judah, & Murphree, 1955; Gilgash, 1957), whereas others reported no changes (Judson & Mac Casland, 1960; Pare-

des, Baumgold, Pugh, & Ragland, 1966). One study (Gilgash, 1957) indicated that a group of schizophrenic patients showed a greater point average increase on the Comprehension subtest than on any other subtest of the Verbal scale subsequent to chlorpromazine medication. Since most of these studies were reported in the late 1950s and similar assessment by the WAIS has been rarely reported in the past decade, it seems that psychometric evaluation has been a neglected resource as a clinical dependent variable in determining the impact of drug treatment on intellectual functioning.

The manner in which the Comprehension subtest and AMPT covaried in the design of the present study suggests that this subtest, or possibly other subtests of the WAIS, in contradistinction to conventional behavioral rating scales, offers some promise for understanding the complexity of interacting variables in cognitive psychopharmacology. Further studies of drug-induced intellectual changes in schizophrenic patients should be directed at determining whether any of the other subtests of the WAIS are as sensitive as Comprehension in assessing changes and whether Comprehension is sensitive to the effects of drugs other than AMPT and L-dopa in designs similar to the present study.

Reference Note

1. Nasrallah, H., Rogol, A., Wyatt, R. J., & Gillin, J. C. Potentiation of phenothiazine-induced prolactin secretion by alpha-methyl-para-tyrosine (AMPT) in schizophrenic males. Unpublished manuscript, National Institute of Mental Health, Washington, D.C., 1977.

References

- Abrams, J. Chlorpromazine in the treatment of chronic schizophrenia. *Diseases of the Nervous System*, 1958, 19, 20-28.
- Barbeau, A. Long-term appraisal of levodopa therapy. *Neurology (Minneapolis)*, 1972, 22, 22-24.
- Cotzias, G. C., Papavasiliou, P. S., & Gellene, R. Modification of parkinsonism: Chronic treatment with L-dopa. *New England Journal of Medicine*, 1969, 280, 337-345.
- Donnelly, E. F., & Chase, T. N. Intellectual and memory function in parkinsonian and non-parkinsonian patients treated with L-dopa. *Diseases of the Nervous System*, 1973, 33, 119-123.
- Engelhardt, D. M., Freedman, N., Rosen, B., Mann, D., & Margolis, R. Phenothiazines in prevention of psychiatric hospitalization: III. Delay or prevention of hospitalization? *Archives of General Psychiatry*, 1964, 11, 162-169.
- Gardner, M. J., Hawkins, H. M., Judah, L. N., & Murphree, O. D. Objective measurement of psychiatric changes produced by chlorpromazine and reserpine in chronic schizophrenia. *Psychiatric Research Reports*, 1955, 1, 77-83.
- Gilgash, C. A. Effects of thorazine on Wechsler scores of adult catatonic schizophrenics. *Psychological Reports*, 1957, 3, 561-564.
- Green, R. A., Bigelow, L., O'Brien, P., Stahl, D., & Wyatt, R. J. The Inpatient Behavioral Rating Scale. A 26-item scale for recording nursing observation of patients' mood and behavior. *Psychological Reports*, 1977, 40, 543-549.
- Judson, A. J., & Mac Casland, B. W. The effects of chlorpromazine on psychological test scores. *Journal of Consulting Psychology*, 1960, 24, 192.
- Lasky, J. J., Klett, C. J., Caffey, E. M., Bennett, J. L., Rosenblum, M. P., & Hollister, L. E. Drug treatment of schizophrenic patients. *Diseases of the Nervous System*, 1962, 23, 698-706.
- Nasrallah, H. A., et al. Inhibition of dopamine synthesis: Clinical ineffectiveness in chronic schizophrenic patients. A double-blind crossover trial with alpha-methyl-para-tyrosine (AMPT). *Archives of General Psychiatry*, 1977, 34, 649-655.
- National Institute of Mental Health—Psychopharmacology Service Center Collaborative Study Group. Phenothiazine treatment in acute schizophrenia: Effectiveness. *Archives of General Psychiatry*, 1964, 10, 246-261.
- Overall, J. E., & Gorham, D. R. The Brief Psychiatric Rating Scale. *Psychological Reports*, 1962, 10, 799-812.
- Paredes, A., Baumgold, J., Pugh, L., & Ragland, R. Clinical judgment in the assessment of psychopharmacological effects. *Journal of Nervous and Mental Disease*, 1966, 142, 153-160.
- Snyder, S. H., Banerjee, S. P., Yamamura, H. I., & Greenberg, D. Drugs, neurotransmitters, and schizophrenia. *Science*, 1974, 184, 1243-1253.
- Spector, S., Sjoerdsma, A., & Udenfriend, S. Blockade of endogenous norepinephrine synthesis by alpha-methyl-tyrosine, an inhibitor of tyrosine hydroxylase. *Journal of Pharmacology and Experimental Therapeutics*, 1965, 147, 86-95.
- Spitzer, R. L., Endicott, J., & Robins, E. Clinical criteria for psychiatric diagnosis and DSM-III. *American Journal of Psychiatry*, 1975, 132, 1187-1192.
- Yahr, M. D., Duvoisin, R. C., Shear, M. J., Barrett, R. E., & Hoehn, M. M. Treatment of parkinsonism with levodopa. *Archives of Neurology*, 1969, 21, 343-354.
- Yaryura-Tobias, J. A., Wolpert, A., Dana, L., & Merlis, S. Action of L-dopa in drug induced extrapyramidalism. *Diseases of the Nervous System*, 1970, 31, 60-63.

Psychophysiological Effects of Progressive Relaxation in Anxiety Neurotic Patients and of Progressive Relaxation and Alpha Feedback in Nonpatients

Paul M. Lehrer

Department of Psychiatry
College of Medicine and Dentistry of New Jersey
Rutgers Medical School

Ten anxiety neurotic patients were given four sessions of individual instruction in progressive relaxation, and 10 patients served as waiting list controls. Ten nonpatients were assigned to each of the same conditions, and an additional 10 nonpatients were given four sessions of alpha feedback. Nonpatients showed more psychophysiological habituation over sessions than patients in response to hearing five very loud tones and to a reaction time task. Patients, however, showed greater physiological response to relaxation than did nonpatients. After relaxation, the autonomic responses of the patients resembled those of the nonpatients. The effects of relaxation were more pronounced in measures of physiological reactivity than in measures of physiological activity. Defensive reflexes yielded to orienting reflexes more readily in nonpatients than in patients. There was also a tendency for progressive relaxation to generalize to autonomic functions more than alpha feedback.

Progressive relaxation is one of the most venerable techniques of behavior therapy and one of the most investigated. Jacobson (1938) reviewed a number of the early case studies showing that progressive relaxation training produces extraordinarily low levels of muscle tension, and that patients suffering from a variety of psychological and psychosomatic disorders experience significant relief when they practice the technique. Gellhorn (1958) hypothesized that progressive relaxation reduces physiological reactivity through reduction in proprioceptive feedback from the muscles to the reticular system. However,

Davison (1966) disputed this "peripheralist" theory and argued that relaxation works on a more central, cognitive level, since subjects can still be anxious when their muscles are rendered almost completely flaccid by curare. A number of controlled studies have been done on the treatment effectiveness of relaxation. Relaxation has been found to be effective in treating the symptoms of insomnia (Borkovec & Fowles, 1973; Borkovec, Kaloupek, & Slama, 1975; Lick & Heffler, 1977; Nicassio & Bootzin, 1974; Woolfolk, Carr-Kaffashan, McNulty, & Lehrer, 1976); hypertension (Deabler, Fidel, Dillenkoffer, & Elder, 1973; Shoemaker & Tasto, 1975); tension headaches (Cox, Freundlich, & Meyer, 1975); and chronic anxiety (Townsend, House, & Addario, 1975).

This research was supported by Biomedical Research Support Grant 27-1985 from Rutgers Medical School.

The author is indebted to Ric Belser and Linda Duberstein, who trained the subjects and scored the data; to Alan Jusko, who got the equipment to work; to John Carpenter for his statistical advice; and to Robert Edelberg for his advice and support in all phases of this study.

Requests for reprints should be sent to Paul Lehrer, Department of Psychiatry, Rutgers Medical School, Piscataway, New Jersey 08854.

The first controlled studies on the psychophysiological effects of relaxation were done on reflexes of the skeletal muscles. Miller (1926) tested the finger withdrawal reflex in subjects who had been trained in progressive relaxation for 6 months. The reflex was smaller in subjects who were instructed to

relax than in those who were instructed not to do so. A more sophisticated control was added by Freeman (1933), who instructed all subjects to relax but had some subjects sit in a seat that required them to maintain some muscle tension in order to remain comfortable. The finger withdrawal reflex was smaller in the relaxation group in this study also. More modern studies of skeletal muscle activity have indicated that progressive relaxation can reduce frontalis electromyogram (EMG) activity more than control conditions (Cox et al., 1975). One study found that tape-recorded instructions in progressive relaxation reduced frontalis EMG activity more than did autogenic training (Staples, Coursey, & Smith, Note 1).

Since progressive relaxation operates primarily on the skeletal muscles, it can be expected to have a greater effect on this system than on other physiological systems. Davidson and Schwartz (1976) have argued that all relaxation techniques are not alike and that they probably have different effects on different physiological systems, depending on the system at which they are most directly aimed. Nevertheless, a number of studies have shown that under certain conditions, autonomic measures are also affected by progressive relaxation training. For example Paul (1969b) found that brief live progressive relaxation training reduced heart rate and respiration rate more than a control condition, and it reduced heart rate more than hypnotic suggestions to relax.

However, results have not been consistent. In another study, Paul (1969a) found that although live progressive relaxation training reduced a combined measure of physiological reactivity to an anxiety-provoking scene, no individual physiological measure was significantly reduced. Lehrer (1972) found that although relaxation reduced heart rate toward the end of a period of electric shock, it had no measurable effect on skin potential or respiration rate. Other studies have found no differences in autonomic activity between progressive relaxation and ordinary rest (Paul & Trimble, 1970; Grossberg, Note 2) or suggestions to relax (Edelman, 1970). Even EMG measures have not always significantly differentiated progressive relaxation training from

control conditions (Haynes, Moseley, & McGowan, 1975; Lehrer, 1972; Paul & Trimble, 1970). A number of factors may account for the discrepant results. Among these are population differences, techniques for independently assessing level of anxiety, intensity of training, and conditions of testing.

Population Differences

The effects of relaxation appear to be clearer among persons with high basal anxiety levels. Thus Wilson and Wilson (1970), studying patients in a general medical hospital, found that brief relaxation instructions reduced heart rate as compared with a control group that listened to a talk on the history of baseball, but only in high anxious subjects, as determined by responses to the IPAT Anxiety Scale. Also, Brandt (1974), studying college students, found that taped progressive relaxation instructions reduced the size and frequency of electrodermal responses, heart rate, and EMG when measured during relaxation training, but when these measures were repeated for a rest period after the second (and last) training session, only heart rate was different between the groups, and only among high scorers on the Fear Survey Schedule. It should come as no surprise that the effects of relaxation are easier to measure among anxious rather than nonanxious persons, since nonanxious persons are presumably able to relax without special training. Lader and Wing (1966) reported that anxiety neurotics do not show habituation of the skin conductance to repeated presentations of a loud (100 dB, 1,000 Hz) aversive tone. They also reviewed a number of other studies showing that anxiety neurotics have higher levels of sympathetic activity, slower habituation of the electrodermal response to noxious stimuli, and higher levels of some indicators of physiological arousal.

Method of Assessment of Anxiety

Although the studies by Brandt (1974) and by Wilson and Wilson (1970) suggest that the IPAT anxiety inventory and the Fear

Survey Schedule hold promise as screening devices for relaxation studies, most of the commonly used paper-and-pencil tests of anxiety probably do not measure the kind of gross psychopathology that was studied by Lader and Wing (1966). For example, Edelman (1970) used the Spielberger State-Trait Anxiety Inventory (STAI) as a screening device and found that neither anxious nor non-anxious subjects manifested any significant autonomic changes after tape-recorded relaxation instruction. Other paper-and-pencil tests may have the same problem, especially the commonly used Taylor Manifest Anxiety Scale (TMAS). In some instances the TMAS has been found to be significantly related to physiological measures, such as eyeblink conditioning (Spence & Taylor, 1966), conditioning of finger withdrawal reflex (Sloane, Davidson, & Payne, 1965), the Palmar Sweat Index during a verbal conditioning task (Haywood & Spielberger, 1966), conditioned heart rate accelerations (Dube, 1966), and basal heart rate (Lehrer, 1969). However, most studies have found no relationship between the TMAS and autonomic measures (e.g., Beam, 1955; Bitterman & Holtzman, 1952; Bursten & Russ, 1965; Calvin, McGuigan, Tyrrell, & Soyars, 1956; Lewinsohn, 1956; McGuigan, Calvin, & Richardson, 1959; Sloane et al., 1965; Spelman, 1966).

A similar lack of significant relationships has also been found between various physiological measures and the Maudsley Neuroticism Scale (Spelman, 1966; Sloane et al., 1965), Welsh's A factor (Katkin, 1965), McReynold's Assimilation Scale (McReynolds, Acker, & Brackbill, 1966), and several tests of situational anxiety (Katkin, 1966; Rosenstein, 1960). A possible reason for the inadequacy of these paper-and-pencil tests as measures of anxiety is that these tests may measure other things besides anxiety, for example, the willingness of people to describe themselves negatively or to admit emotional weaknesses. Thus Kimble and Posnick (1967) reported a correlation of .73 between the TMAS and a scale having nothing to do with anxiety that contained items matched with the TMAS on emotional importance and social acceptability.

Intensity of Training

None of the studies using brief relaxation instructions produced the size of changes implied in Jacobson's work (1938), in which people receive training over a period of months. Perhaps the intensive training given by Jacobson and his more strict adherents has a more profound physiological effect than the brief training procedures that are conventionally used, especially those that are administered in a single session and/or are given by tape recording.

Conditions of Testing

In almost all of the above cited studies that found brief relaxation training to be effective in altering psychophysiology, subjects were tested either while they were receiving relaxation instruction or while the trainer was still in the room immediately after live instruction. These conditions are not analogous to the clinical situation in which the patient must apply the training at home and may have exaggerated the effects of the relaxation training. This is suggested by Paul and Trimble's (1970) findings that live relaxation training (with the therapist in the room during testing) produces a greater physiological effect than tape-recorded instructions, and by Brandt's (1974) findings that the physiological effects of brief tape-recorded relaxation instructions are quickly attenuated after a training session is over.

Alpha Feedback

Electroencephalogram (EEG) alpha control has recently become a topic of broad interest and wide application. Early investigators have remarked on how the presence of alpha appears to be inversely related to anxiety (Jasper, 1937), and some modern work on alpha feedback has found that among subjects who do learn to control their alpha production, the production of alpha is usually associated with subjective feelings of relaxation, letting go, and a lack of focused thought (Brown, 1970; Nowlis & Kamiya, 1970). More recently, however, Orne and Paskewitz (1974) reported a study showing that under the

threat of shock, subjects who are given occipital alpha feedback still show increases in subjective anxiety, frequency of skin conductance responses, and heart rate, despite the fact that their levels of occipital alpha production remain high. Thus, although high levels of alpha are, under some conditions, associated with diminished anxiety, the production of alpha appears to be at least partially a differentiable process from subjective anxiety and autonomic reactivity.

Present Study

This study compared the physiological effects of progressive relaxation, alpha feedback, and a no-treatment control condition. The responses of psychiatric patients who were clinically diagnosed as suffering from severe anxiety were compared with that of non-anxious volunteer subjects who were not psychiatric patients. Subjects received approximately the same intensity of training as is usually given to patients in behavior therapy, and the testing was done during a separate session by an experimenter who was not involved in training the subjects to relax.

Loud Tones

The testing situation included a series of loud tones similar to those used by Lader and Wing (1966), who had found that tranquilizing medication allowed anxiety neurotics' skin conductance responses to the tones to habituate at a rate equal to the rate in normal subjects.

Reaction Time Task

The question is frequently asked whether relaxation makes people somnolent and less reactive to the environment in general or whether, as suggested by Wilson and Wilson's (1970) data, it reduces the specific effects of anxiety and *enhances* the ability to pay attention to the environment and to solve problems. A test of reaction times was thus included to test whether relaxation facilitates this type of "alertness" response and whether, on a physiological level, it affects the orienting reflex (cf. Graham & Clifton, 1966).

Alpha Feedback

The present study also included an alpha feedback condition that used a training procedure similar to that described by Brown (1970). The clinical and experimental studies cited above suggest that progressive relaxation would, especially among the anxious patients, generalize to autonomic reactivity, whereas alpha feedback would not generalize as much. Because of the difficulty in finding a sufficient number of suitable anxious patients who would be willing to participate in the study, I decided to test the alpha condition only on nonpatient volunteers. Light was used as a feedback signal rather than the more conventional tone signal so as not to unintentionally habituate these subjects to the noxious tones used in the testing situation. Although light may depress occipital alpha, it has also been found that the presence of light in a room may increase the effects of alpha feedback training (Paskewitz & Orne, 1973).

Method

Subjects. The subjects for this experiment were 20 anxiety neurotic patients and 34 nonpatient volunteers. The patients were recruited through their therapists, who are professionals at the Rutgers Mental Health Center or members of the faculty of the Department of Psychiatry of Rutgers Medical School and/or of the Graduate School of Applied and Professional Psychology at Rutgers University. Patients were accepted whose symptoms were "primarily" those of anxiety (according to the report of their therapists), who were not psychotic, and who were not regularly taking any medication other than minor tranquilizers. Most of the anxious subjects were diagnosed as anxiety neurotic. Patients with histories of neurological or cardiovascular disorders were excluded, and testing sessions were held only after patients had not taken any medication for at least 24 hours. Ten patients each were included in the relaxation and control (ordinary rest) conditions. The nonpatients were recruited from newspaper advertisements, notices to staff members at Rutgers Mental Health Center, and signs on bulletin boards. Subjects were accepted for the experiment only if they reported that they were in excellent physical and emotional health, were not taking any psychoactive medication, and were not currently receiving any form of treatment for an emotional problem. Subjects were assigned to groups on as random a basis as possible, given difficulties in scheduling. Eleven subjects were assigned to the relaxation group, 11 to the alpha feedback group, and 12 to the con-

Table 1
Design of the Experiment

Condition	Population	Pretest	Training
Relaxation	Patients and nonpatients	5-min rest; 5 tones, 100 dB, 1,000 Hz	4-5 sessions of training in progressive relaxation during a 3-week period
Control	Patients and nonpatients	5 reaction time trials with variable interval foreperiods	3-week wait for training
Alpha feedback	Nonpatients	3-min rest	4-5 sessions of alpha feedback during a 3-week period

Note. Posttest consisted of the same procedures as the pretest.

trol group (ordinary rest). Both patients and nonpatients in the control group were offered relaxation training after their participation in the two testing sessions. They were not told that they were members of a control group. The complete design is summarized in Table 1.

Testing procedure. Testing sessions were held approximately 3 weeks apart. Subjects in the relaxation and alpha feedback groups received their training between testing sessions, whereas subjects in the control group were tested at the same intervals but were not in contact with the project between testing sessions.

In each testing session, subjects were first wired to the equipment, and, in both conditions, before and after training, they were asked to close their eyes and to relax as deeply as possible in the reclining chair. They were told to expect to hear some loud tones after several minutes but to try to remain relaxed through them.

After approximately 5 min of resting, five aversive tones were presented for a duration of 1 sec each with an interstimulus interval varying between 30 and 60 sec. They were presented from a speaker at a located several feet behind the subject's chair, at a frequency of 1,000 Hz and an intensity of 100 dB(A).

Approximately 30 sec after the last aversive tone, the experimenter reentered the subject room, put a pushbutton switch in the subject's nondominant hand, and told the subject that the next task consisted of a series of reaction time trials. Subjects were told to expect to hear "softer, deeper" tones than they had heard before and that they would not be aversive. Each tone was a "get ready" stimulus. Subjects were told to press the button as quickly as possible after the tone went off. The tones were presented at 60 dB and approximately 300 Hz for durations varying between 5 and 15 sec, with interstimulus intervals of between 30 and 60 sec. Following the last trial, subjects were given an additional 3-min rest.

Subjects were also administered the State Anxiety scale of the STAI before and after each session. Before each session they were asked to fill it out

with reference to how they felt "right now." After the session they were asked to fill it out as to how they felt "during the session." All subjects were also administered the STAI Trait scale prior to their first session.

Apparatus. Physiological measures for the first 10 subjects tested, approximately equally distributed among groups, were taken on a Grass Model 5 polygraph. The rest of the subjects were tested on a Beckman Type R Dynograph. The following measures were taken: skin conductance, heart rate, and EEG from the dominant occipital area.

Skin conductance was measured from a pair of silver-silver chloride electrodes, 1 cm² each, mounted, in a special holder to facilitate recording from the palmar surface of a finger (Edelberg, Note 3). The middle finger of the dominant hand was used. K-Y jelly was used as the electrolyte. Resistance was measured in kilohms on the Grass polygraph, which was later converted to micromhos conductance for data analysis. On the Beckman polygraph, a Lykken coupler was used for direct measurement of skin conductance. Heart rate was measured through a cardi tachometer from standard electrocardiogram electrodes attached to the right arm and left leg. EEG leads were attached to the dominant occipital area (O₁ or O₂) and to both earlobes as the reference point. A ground lead was attached to the forehead. Alpha was filtered by an instrument, similar to that described by Paskewitz (1971), which provides a voltage output when a criterion amplitude of filtered alpha is present. For the testing sessions, the criterion level for alpha for all was set at 19 μ V.

Relaxation training. Subjects in the progressive relaxation group were administered four sessions of individual training using an abbreviated form of Jacobson's (1938, 1964) technique. Subjects were asked, alternately, to tense and to relax each of 35 muscle groups. Subjects were instructed to tense each muscle as little as necessary to barely feel the proprioceptive sensations of muscle contraction. Each muscle group was worked with repeatedly until the subject could feel the muscle contraction without engaging in any overt movement. Each subject was

trained at his or her own pace. In most subjects the arms and some of the leg muscles were trained during the first session, the remainder of the leg muscles and the trunk were trained during the second session, and the facial muscles were trained during the third session. The fourth session was generally devoted to review and general relaxation. Subjects were asked to rate their relaxation during each session from 0, "as relaxed as I ever felt," to 100, "very anxious." If a subject's self-rating of relaxation did not fall below 10 by the fourth session, an additional session of training was given. Subjects were told to practice the technique at home for 1 hour daily, but they were given no written or taped material to take home with them.

Alpha feedback. Dominant occipital alpha feedback was given in four sessions of individual training. The occipital area was chosen for feedback because of the high density of alpha often recorded from that site and because of previous reports of success in conditioning alpha from that region, including one study showing that increased alpha can persist during aversive stimulation (Chisholm, De Good, & Hartz, 1977). Dominance was assessed simply by asking subjects whether they are right-handed or left-handed. In each session the subjects were first wired to the equipment and were then asked to lie back on a reclining chair, to close their eyes, and to try to keep their minds blank. This, they were told, is the best way to stay in an "alpha state." Through their closed eyelids, they were told, they would see a light flashing on and off. The presence of this light indicated that the subject was emitting alpha. A small flashlight bulb was mounted on the 8-foot (2.4 m) ceiling directly above the subject's head, and it flashed on when the subject's alpha amplitude reached a specified criterion. Generally this criterion was 19 μ V. If the subject emitted less than approximately 30% alpha, the criterion was lowered and the response was shaped. Subjects were given approximately 40 minutes of training during each session. They were told to try to put themselves into an alpha state at home in daily 1-hour practice sessions, without feedback. As in the relaxation group, alpha feedback subjects were asked to rate their level of relaxation after each session and were given a fifth session of training if they did not rate themselves below 10 after the fourth session.

Results

Most analyses reported below are on difference scores between values on the pretest session and values on the posttest session. Each data point in the first session was subtracted from the corresponding point in the second session (posttest). Thus, in the analyses of variance that were computed, the *F* and *p* values for each main effect or interaction are identical to the interaction between that effect and an implicit sessions factor

from an analysis done on the raw session scores (vs. the session-difference scores used here). Covariance adjustments were also used in these analyses to adjust for some differences between experimental groups in the first session that were determined by analyses of variance done on the pretest session alone. The session-difference scores were adjusted for pretest (Session 1) levels of the same measure. Also, to test the significance of habituation across sessions and across trials, analyses of variance were computed on the raw scores for all data points within each of the sessions, and analyses of variance were computed with an additional between-groups measure for session. All analyses include a between-groups factor (treatment) and a within-group factor (time). The time factor consists of five trials each for the loud tones and reaction time task and eight epochs or time samples of 15 sec each for the two rest periods at the beginning and end of each session (i.e., before the first loud tone and after the last reaction time trial). Data were scored for 15 sec before and after each loud tone and for the duration of each get ready stimulus and an equivalent time period before each reaction time trial.

Wherever data points were missing due to equipment difficulties, they were filled in using the unweighted means solution. Data were analyzed on an IBM 370 computer using the Data-Text program (Armor & Couch, 1972) Versions 3.0 and 3.1. Each measure was first tested for skewness, kurtosis, and heteroscedasticity. Where necessary, normalizing transformations were used, and, in all cases, the assumptions of the analysis of variance and covariance were met.

Combined analyses of patients and nonpatients in the relaxation and control conditions. To compare diagnostic groups and treatment effects in a single analysis, the alpha feedback condition, which was only given to nonpatients, was excluded in some analyses. In these analyses the covariance - adjusted session-difference scores have two between-groups measures, diagnosis (patient vs. nonpatients) and treatment (relaxation vs. control), and one repeated measure, time.

Separate analyses of patients and nonpatients. To test the significance of treatment

Table 2
Summary of Pretest and Posttest Means that are Significantly Different Between Groups

Measure	Nonpatients						Patients			
	Relaxation		Control		Alpha		Relaxation		Control	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
State anxiety	42.5	34.8	32.3	31.2	43.0	40.7	52.0	38.5	49.0	44.4
Cardiac decelerations ^a										
Tones	-.5	2.5	.1	.7	.5	1.6	-.8	.3	1.1	.3
RT trials	-.3	2.3	.6	-.4	.0	1.1	1.6	2.1	.6	.2
Cardiac accelerations ^a										
Tones	.2	2.7	4.8	4.8	1.2	.8	3.4	.7	1.6	5.6
RT trials	.9	2.6	.6	1.4	.6	-.8	-.9	-1.4	-.7	.6
Percent alpha										
Initial rest	38.9	27.0	29.1	29.7	12.2	23.8	32.6	41.6	33.7	30.8
Before tones	41.1	30.7	24.8	24.9	16.4	24.8	30.4	46.5	36.9	34.0
Posttones	34.5	28.5	17.7	24.4	9.3	20.9	21.4	42.8	30.9	26.8
Before RT trials	40.9	30.9	27.4	24.4	20.8	22.1	34.1	47.2	38.9	35.6
During RT trials	27.4	33.0	19.0	28.9	9.1	19.0	30.9	41.4	31.8	31.0
Maximum log skin conductance										
Initial rest	2.0	1.5	1.5	1.3	1.5	1.3	2.2	1.5	1.8	2.0
End-of-session rest	2.3	1.7	1.8	1.6	1.8	1.5	2.4	1.4	1.9	2.2
Posttones	2.4	1.9	2.0	1.8	2.1	1.8	2.5	1.8	2.3	2.5
During RT trials	2.5	1.9	2.0	1.9	2.1	1.6	2.5	1.7	2.1	2.3

Note. RT = reaction time.

^a A negative cardiac deceleration reflects a lower minimum heart rate before the stimulus than after the stimulus. A negative cardiac acceleration reflects a higher maximum heart rate before the stimulus than after the stimulus.

and habituation effects in each of the two populations that were studied, the data from each population were submitted separately to the analyses of variance described above, without the between-groups factor of diagnosis.

Subjective Anxiety

Trait Anxiety

As a check on our diagnostic procedures and to check on the randomness of assignment to treatment conditions, a two-factor analysis of variance was done on the STAI Trait Anxiety scale scores, with diagnosis and treatment as between-groups measures. As expected, patients scored significantly ($p < .005$) higher than nonpatients. The treatment effect was not significant, indicating that on this measure, the assignment of subjects to groups was effectively random.

State Anxiety

The STAI State Anxiety scale was analyzed by an analysis of variance with two between-

groups measures (diagnosis and treatment) and two repeated measures (time and session). The two levels of the time factor were pre-session and during session. The Diagnosis \times Time interaction was significant at $p < .004$, indicating that the patients reported themselves to be more anxious than the nonpatients at the beginning of each session. The analysis of variance on Session 1 scores yielded a significant ($p < .02$) effect for treatment, indicating that on this measure, the method of random selection to groups was not adequate and that the subjects assigned to the control group admitted less anxiety than the subjects assigned to the treatment groups (Table 2). Thus an analysis of variance was done on the covariance-adjusted difference scores, as described above. Reported anxiety decreased significantly more in the relaxation group than in the control groups. Separate analyses of the patient and non-patient groups revealed that patients and nonpatients both habituated over sessions but

Figure 1. Differences between the pretest and the posttest, adjusted for the level for each score during the pretest session.

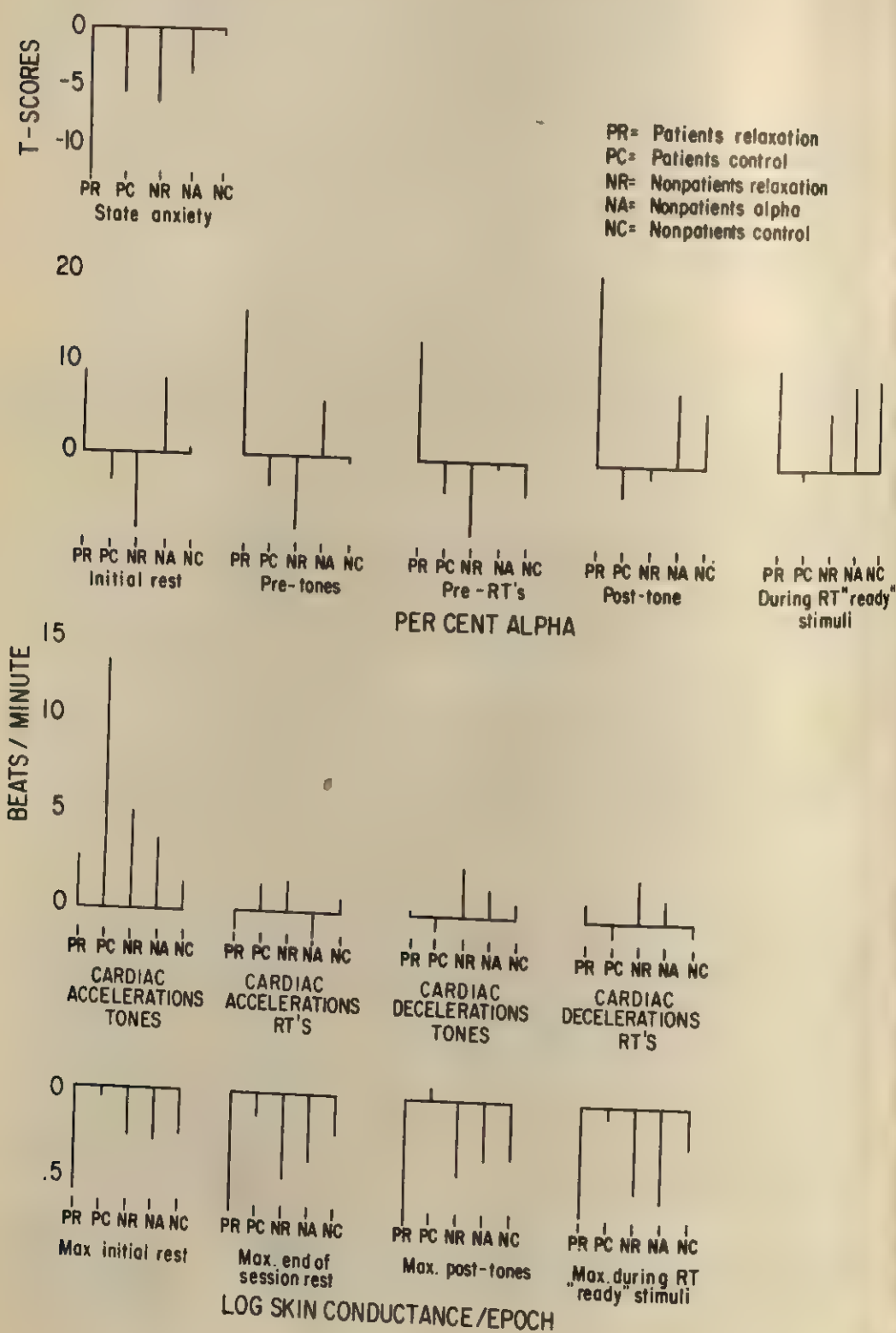


Table 3
Summary of Effects

Measure	Habituation over trials			Habituation over sessions			Treatment effect ^a		
	Patients	Non-patients	Com-bined ^b	Patients	Non-patients	Com-bined ^b	Patients	Non-patients	Com-bined ^b
State anxiety ^c	**		*****	**	*****	*****		*	**
Percent alpha ^d									
Initial rest	—	—	—						***
Before tone	—	—	—				*		***
After tone	*	*****	*****	***			**		***
Before each RT trial	—	—	—				*		***
During RT trial	**	*****	*****		*****	*			*
Alpha blocking									
Tones		*****	*		*	*			
RT trials		*	*****		*****	*****			
Heart rate acceleration									
Tones	*****	*****	*****				**		***
Heart rate deceleration ^f									
Tones					*****				
RT trials							*	*	*****
Log skin conductance									
Initial rest	—	—	—		*	*			*
End of session rest	—	—	—		*	*			*
Tone onset	—	—	—		*	*	*		*
RT onset	—	—	—		*****	*****			*
Maximum skin condition/epoch									*
Initial rest	—	—	—		*	*			*
End of session rest	—	—	—		*	*			*
Posttone	**	*****	*****	*	*****	*****	**		**
During RT trials	*****	*****	*****		*****	*****	***		**
Amplitude of SCR									
Tone ^g	*****	*****	*****			***			*
RT trials ^h	*****	*****	*****			*			
SCR recovery rate									
Tones		*****	*****						

Note. RT = reaction time.

^a Unless otherwise specified, this refers to the treatment effect on the covariance-adjusted session-difference scores.

^b This column refers to an analysis of both the patients and the nonpatients, for the relaxation and control conditions only.

^c Patients emitted higher skin conductance responses (SCRs) to the first tone than nonpatients and rated themselves as more anxious than nonpatients at the beginning of each session. The Diagnosis \times Time interactions are, respectively, significant at $p < .03$ and $p < .001$.

^d Increases in alpha are scored as habituation.

^e These values are for the Treatment \times Diagnosis effect on the covariance-adjusted session-difference scores. The overall Treatment effect is not significant.

^f Here an increase in cardiac deceleration is rated as habituation.

^g The Session \times Trial effect is also significant at $p < .05$, indicating that habituation over trials is faster in the posttest than in the pretest.

^h Patients also emitted significantly smaller skin conductance responses than nonpatients.

* $p < .10$.

** $p < .05$.

*** $p < .01$.

**** $p < .005$.

that the covariance-adjusted treatment effects were significant only for the nonpatients. A trend analysis on the session-difference scores, with the nonpatient relaxation group assigned a value of 1, the alpha feedback group a value of 2, and the nonpatient control group a value of 3, revealed a linear trend of borderline significance ($p < .052$), suggesting that relaxation was most effective in reducing anxiety, the alpha feedback condition second, and the control condition least (Figure 1).

EEG Alpha

Percent alpha

Percent alpha of 19 μ V or greater was scored for each of the time periods described above. As can be seen in Figure 1, relaxation appears to have increased patients' tendencies to emit alpha waves. However, paradoxically, percent alpha appears to have *decreased* over time in the nonpatient relaxation group relative to the nonpatient alpha and control groups. The Diagnosis \times Treatment effect was significant for most time periods (Table 3) in the combined analysis. Analyses of the pretest session scores revealed significant differences between treatment groups ($p < .02$) for the epochs following the loud tones, with

the alpha feedback group having less alpha than the other two groups (Table 2). Covariance-adjusted analyses of variance on the pretest versus posttest change scores revealed that among the patients, differences between treatments were significant for the epochs following the loud tones and were of borderline significance for the epochs preceding the loud tones and reaction time trials. For the nonpatients, none of the differences were significant. A trend analysis was computed for the nonpatients across groups on the session-difference scores, with the relaxation group assigned a value of 1, the alpha feedback group a value of 2, and the control group a value of 3. A significant ($p < .05$) quadratic trend was found for the beginning-of-session rest, before and after the loud tones, and before the reaction time trials, thus indicating that the alpha feedback condition produces greater increases in percent alpha than the other two groups of nonpatients.

Alpha Blocking

Alpha blocking was assessed by subtracting the prestimulus percent alpha from the post-stimulus value for each trial of the loud tones and the reaction time "ready" stimuli. There were no differences on this measure as a func-

Table 4

Differences Between Patients and Nonpatients in Heart Rate

Measure	<i>M</i>		<i>F</i>	<i>p</i>
	Patients	Nonpatients		
HR at stimulus onset				
Tones	77.7	71.2	2.50 ^a	.12
RT trials	76.4	70.4	2.10 ^a	.16
Minimum HRs				
Pretone	72.3	64.6	3.71 ^a	.06
Posttone	72.4	64.0	4.52 ^a	.04
Pre-RT trials	72.2	64.7	3.73 ^b	.06
During RT stimuli	71.2	63.7	3.33 ^a	.08
Maximum HRs				
Pretone	83.2	76.7	2.42 ^a	.13
Posttone	86.1	80.2	2.08 ^a	.16
Pre-RT trials	81.2	76.2	1.46 ^b	.24
During RT stimuli	80.7	77.3	.63 ^b	.43

Note. Discrepancies in degrees of freedom are due to missing data. HR = heart rate; RT = reaction time.

^a $df = 1, 36$.

^b $df = 1, 35$.

^c $df = 1, 33$.

tion of either diagnosis or treatment. Although the combined analysis showed significant main effects for sessions and for trials, the separate analyses on the two populations indicate that this effect was only significant for the nonpatients. Patients did not habituate over trials on this measure, but nonpatients did. The Diagnosis \times Session interaction, which reflects the difference between habituation rates among the two populations in the combined analysis, was significant at $p < .005$ for the reaction time "get ready" stimuli.

Heart Rate

Tonic heart rate was measured every 15 sec during the rest periods and immediately before each trial during the loud tones and the reaction time task. There were no significant effects for treatment. However, the main effect for diagnosis was significant for *minimum* heart rates in the epochs following the loud tones and approached significance ($p < .07$ and $p < .08$) for minimum heart rates at other times (Table 4). This indicates that patients had higher "floor" heart rates than nonpatients. Although the pattern was similar for *maximum* heart rates in each epoch, the p values were much higher ($p < .2-.5$), thus strongly suggesting that peak heart rates were unaffected by anxiety neurosis.

Heart rate accelerations to the tones and reaction time stimuli were assessed by subtracting the maximum prestimulus heart rate from the maximum poststimulus heart rate. This adjustment was used to control for the effects of sinus arrhythmia, which otherwise would obscure the effects of reactivity to the tones. It was used rather than the traditional calculation of the difference between heart rate at stimulus onset and poststimulus maximum or minimum levels (uncorrected for prestimulus variability) because of experience from a prior study on the effects of alcohol on cardiac reactivity (Lehrer & Taylor, 1974). That study also used loud 1-sec long noxious tones. In Lehrer and Taylor (1974), the heart rate accelerations and decelerations corrected for prestimulus maximum and minimum levels differentiated between drunk and sober conditions, whereas the uncorrected accelerations and decelerations did not. Drunk subjects

showed greater corrected decelerations and smaller corrected accelerations in response to the five loud tones. Cardiac accelerations may be interpreted as components of a defensive reflex, and cardiac decelerations, as components of an orienting reflex to the stimuli (Graham & Clifton, 1966).

Accelerations

The differences between treatment groups in the first session (Table 2) were not significant. Cardiac accelerations elicited by the loud tones habituated significantly over trials (Table 3). Also, relaxation significantly enhanced habituation of this reflex over sessions, but only among the patient groups (see Tables 2 and 3 and Figure 1). There were no significant differences between groups for cardiac accelerations to the reaction time task.

Decelerations

The differences between treatment groups in the first session (Table 2) were not significant. Cardiac decelerations increased over sessions among the nonpatients but not among the patients. Significant treatment effects on the covariance-adjusted session-difference scores were obtained for the reaction time trials in the combined analysis and among the nonpatients. This effect was of borderline significance for the patients. Assuming that progressive relaxation would produce greater generalization to heart rate than alpha feedback, a linear trend analysis was done between groups of nonpatients, with the relaxation group assigned a value of 1, the alpha feedback group a value of 2, and the control group a value of 3. Significant linear trends were found for the tones ($p < .05$) and for the reaction time trials ($p < .02$), indicating that this effect was greater in the relaxation group, least in the control group, and at an intermediate level in the alpha feedback group.

Skin Conductance

Tonic Levels

Skin conductance was measured every 15 sec during the rest periods and immediately

before each trial during the loud tones and reaction time task. At each of the periods studied, skin conductance appeared to decrease over sessions in the relaxation group and to increase in the control group. The effects appear to be more pronounced among the patients than among the nonpatients, but they are at best, of only borderline significance (Table 3). Overall habituation across sessions was significant for the nonpatients but not for the patients.

Reactivity

The maximum skin conductance level during each epoch was scored and submitted to analysis. The differences between treatment groups (Table 2) in the first session were significant ($p < .04$) for the reaction time ready stimuli and the end-of-session rest periods in the combined analysis of relaxation and no treatment for patients and nonpatients. They were not, however, significant in either analysis of the diagnostic groups taken separately. The pattern of results was the same as that for tonic levels of skin conductance, but this time the covariance-adjusted treatment effects were significant for the patients (Tables 2 and 3, Figure 1). These results indicate that although the nonpatients tended to habituate over sessions regardless of treatment, the patients only habituated when treated. All subjects, however, significantly habituated over trials (within sessions) to the tones and reaction time task. The tendency of this measure to decrease more after relaxation instructions than after alpha feedback was only of borderline significance. The linear trend across groups, computed as described above for cardiac accelerations, approached significance during the end-of-session rest ($p < .13$) and during the reaction time trials ($p < .07$).

Frequency of skin conductance responses was also analyzed. During the rest periods at the beginning and end of each session, skin conductance increases of greater than 1% of the prestimulus skin conductance were tallied. The reciprocal transformation rendered these scores normally distributed. There were no significant effects on this measure.

Another measure of electrodermal reactivity

that was studied is the amplitude of the skin conductance responses elicited by the loud tones and the reaction time task. This was defined as the difference between the skin conductance at stimulus onset and the maximum skin conductance reached during the 15 sec following each loud tone or during the get ready stimulus for each reaction time trial. A log transformation rendered this measure normally distributed. The skin conductance response habituated significantly over sessions and over trials for the loud tones and the reaction time task. Also, a significant ($p < .05$) Session \times Trials interaction revealed that subjects habituated more quickly in the second testing session than in the first during the loud tones. The patients made a larger response to the first loud tone than the nonpatients, but otherwise the responses to the tones were similar for the two groups. The Diagnosis \times Trial interaction was significant at $p < .02$. On the other hand, the patients emitted significantly ($p < .02$) smaller responses to the reaction time task than the nonpatients. There were no significant effects or interactions involving the treatment factor.

A third measure of electrodermal reactivity was recovery rate of the skin conductance response. Edelberg (1972) has found that stress and threat of stress slow the recovery of the skin conductance response and that schizophrenic patients show slower recovery than normal subjects (Maricq & Edelberg, 1975). He has also demonstrated that the recovery rate can be approximated by measuring the percent recovery during the 2 sec following the apex of each response (Edelberg, Note 3). We computed this measure for the first observable skin conductance response emitted after each loud tone and after the onset of each get ready stimulus for the reaction time task. The log transformation rendered these data normally distributed. Log percent recovery increased significantly over trials during the loud tones. This parallels the findings for cardiac accelerations and appears to reflect a habituation of the defensive reflex over trials. However, there were no significant effects involving diagnosis, treatment, or session.

Reaction Times

The natural log transformation rendered the reaction time data normally distributed. There were no significant effects for treatment, diagnosis, or the interactions of these with sessions.

Discussion

Relaxation Effects

Relaxation appeared to reduce physiological reactivity. This finding was most pronounced among the patients, for whom relaxation was significant for maximum skin conductance elicited by both tasks, for cardiac accelerations (defensive reflexes) elicited by the loud tones, and for percent alpha levels after the loud tones. The effect was not present for measures of tonic physiological activity. This supports the notion that progressive relaxation of the muscles generalizes to relaxation in other physiological systems. It also confirms the findings reviewed above that physiological effects of brief relaxation instructions can best be measured in an anxious population. The reason for this appears to be different for each of the three measures on which significant effects were found. For maximum skin conductance after the stimuli, the nonpatients all appeared to habituate over sessions, whereas the patients only habituated when they had been taught relaxation. For cardiac accelerations to the loud tones, subjects tended to sensitize over sessions rather than to habituate, although this overall effect was not significant. The patient population appears to have been more prone to this sensitization effect than the nonpatients, and this effect was attenuated only by relaxation training. After training, the patients' pattern of autonomic reactivity became similar to that of the nonpatients. For percent alpha after the loud tones, the results are more complex. Percent alpha was significantly increased by relaxation among the patients. However, the nonpatient relaxation group had a seemingly paradoxical decrease in percent alpha, rather than an increase. One can only speculate on possible causes for this. The patient control group showed this same pattern, and the latter

group apparently sensitized over sessions. It is possible that something similar happened to the nonpatient relaxation group, but this would be inconsistent with the data on autonomic functions. Perhaps the nonpatient relaxation group was emitting theta activity. The tape-recorded data were no longer available for checking this, but a cursory examination of the raw data indicate that this was not so. Another possibility is that the nonpatients were "working too hard" at relaxing and were concentrating on their muscles so much that alpha was blocked. I have informally tried concentrating on relaxing my own muscles while being given alpha feedback and have similarly experienced a decrease in alpha production. But why, then, did this not happen with the patient relaxation group? Here we must speculate some more. Perhaps the patients had taken the task more seriously and had practiced more intensively at home. Thus, when they came for testing, the task could have been more automatic for the patients than for the nonpatients. Informal contacts with the subjects indicated that patients were more motivated to practice the task than nonpatients, since they expected some therapeutic benefit.

There are two measures on which the relaxation effects were somewhat more pronounced among nonpatients than among patients: cardiac decelerations (orienting reflexes) to the reaction time task and state anxiety. The slightly greater treatment effects among nonpatients for cardiac decelerations may partly be a mathematical artifact of changes in cardiac accelerations and partly a reflection of the higher heart rate "floor" for the patients, perhaps due to their higher state of general sympathetic tuning. Relaxation apparently did not affect heart rate floors. In a previous study we found that alcoholics also have higher floor heart rates than "normal" subjects (Lehrer & Taylor, 1974), and we interpreted this as reflecting changes in cardiac function due to chronic alcohol intake. It is possible that the effect in the present study was also due to drug intake in the patient group, since most of the patients in the sample had histories of taking tranquilizers. It is also possible that the results of both

studies are due to the effects of chronic anxiety in both patient groups. Perhaps chronic anxiety inhibits the individual from emitting orienting reflexes to situations that had previously evoked defensive reflexes. This "shift" would be expected to occur over time among normal individuals (Sokolov, 1963). In this regard, note also that cardiac decelerations (orienting reflexes) to the loud tones increased significantly over sessions in the nonpatients but not in the anxious patients (Table 3).

The fact that state anxiety was decreased by relaxation more among the nonpatients than among the patients may result from the fact that many of the patients showed obvious evidence of "secondary gain" from their symptoms, including having a friend, parent, or spouse accompany them to each session because of their fears. To admit that their fear was reduced during the testing sessions would have justified their giving up this dependence.

Relaxation, Alpha Feedback, and Generalization

As revealed in the trend analyses on the nonpatients, alpha feedback produced greater increases in alpha but smaller changes in other physiological measures and subjective anxiety than progressive relaxation. These findings give some support to the notion of physiological specificity of alpha. These results are also consistent with the findings of Orne and Paskewitz (1974) that increases in alpha produced by alpha feedback do not generalize well to other physiological functions or to subjective anxiety. Muscle relaxation appears to produce greater generalization.

Physiological Activity Versus Reactivity

It should be noted that relaxation effects were more pronounced for measures of physiological reactivity than for measures of tonic physiological activity. They were greater for maximum skin conductance level for each epoch than for initial skin conductance level in the epoch, greater for cardiac accelerations and decelerations than for tonic heart rate, and greater for percent alpha in the pe-

riod following the loud tones than at any other time period. In terms of Routtenberg's (1968) hypothesis of two separate arousal systems, perhaps progressive relaxation has greater effects of the type of arousal mediated by the limbic system than that mediated by the reticular system.

Other Differences Between Patients and Nonpatients

Another finding of interest is that the patients appeared to be hyperresponsive to threat but somewhat underresponsive to a nonthreatening task, which generally produces an orienting reflex rather than a defensive reflex. The patients emitted larger skin conductance responses to the first loud tone and higher scores on the STAI at the beginning of each session than the nonpatients, but the patients emitted smaller skin conductance response amplitudes during the reaction time task than the nonpatients. Apparently, patients respond with greater extremes of electrodermal reactivity than do the nonpatients, giving smaller responses to the nonthreatening reaction time task (which requires orientation to the environment rather than defense from it) but a larger response to the threatening first loud tone. This squares with the clinical picture of anxiety neurotics as easily upset by everyday stress and not always as "tuned in" to their environment because of their anxious preoccupations.

Finally, Lader and Wing's (1966) contention that people suffering from anxiety neurosis habituate more slowly than nonpatients to noxious stimuli is lent some support by this experiment. Patients showed less habituation of alpha blocking and skin conductance levels than did nonpatients.

Reference Notes

1. Staples, R., Coursey, R., & Smith, B. *A comparison of EMG feedback, autogenic and progressive training as relaxation techniques*. Paper presented at the meeting of the Biofeedback Research Society, Monterey, California, October 1975.
2. Grossberg, J. M. *The physiological effectiveness of brief training in differential muscle relaxation* (Tech. Rep. No. 9). La Jolla, Calif.: Western Behavioral Sciences Institute, 1965.

3. Edelberg, R. Personal communication, August 30, 1973.

References

- Armor, D. J., & Couch, A. S. *Data-text primer*. New York: Free Press, 1972.
- Beam, J. C. Serial learning and conditioning under real life stress. *Journal of Abnormal and Social Psychology*, 1955, 51, 543-555.
- Bitterman, M. E., & Holtzman, W. H. Conditioning and extinction of the galvanic skin response as a function of anxiety. *Journal of Abnormal and Social Psychology*, 1952, 47, 615-623.
- Borkovec, T. D., & Fowles, D. C. Controlled investigation of the effects of progressive and hypnotic relaxation on insomnia. *Journal of Abnormal Psychology*, 1973, 82, 153-158.
- Borkovec, T. D., Kaloupek, D. G., & Slama, K. M. The facilitative effect of muscle tension-release in the relaxation treatment of sleep disturbance. *Behavior Therapy*, 1975, 6, 301-309.
- Brandt, K. The effects of relaxation training with analog HR feedback on basal levels of arousal and response to aversive tones in groups selected according to Fear Survey Schedule scores. *Psychophysiology*, 1974, 11, 242. (Abstract)
- Brown, B. B. Recognition of aspects of consciousness through association with EEG alpha activity represented by a light signal. *Psychophysiology*, 1970, 6, 442-452.
- Bursten, B., & Russ, J. J. Preoperative psychological state and corticosteroid levels of surgical patients. *Psychosomatic Medicine*, 1965, 27, 309-316.
- Calvin, A. D., McGuigan, F. J., Tyrrell, S., & Soyars, M. Manifest anxiety and the Palmar Sweat Index. *Journal of Consulting and Clinical Psychology*, 1956, 20, 356.
- Chisholm, R. C., De Good, D. E., & Hartz, M. A. Effects of alpha feedback training on occipital EEG, heart rate, and experiential reactivity to a laboratory stressor. *Psychophysiology*, 1977, 14, 157-163.
- Cox, D. J., Freundlich, A., & Meyer, R. G. Differential effectiveness of feedback, verbal relaxation instructions, and medication placebo with tension headaches. *Journal of Consulting and Clinical Psychology*, 1975, 43, 892-898.
- Davidson, R. J., & Schwartz, G. E. The psychobiology of relaxation and related states: A multi-process theory. In D. Mostofsky (Ed.), *Behavior control and modification of physiological activity*. New York: Prentice-Hall, 1976.
- Davison, G. Anxiety under total curarization: Implications for the role of muscular relaxation in the desensitization of neurotic fears. *Journal of Nervous and Mental Disease*, 1966, 143, 443-448.
- Deabler, H. L., Fidel, E., Dillenkoffer, R. L., & Elder, S. T. The use of relaxation and hypnosis in lowering high blood pressure. *American Journal of Clinical Hypnosis*, 1973, 16, 421-425.
- Dube, C. S. Conditioned heart rate as a function of manifest anxiety and drug action. *Psychological Reports*, 1966, 19, 1251-1255.
- Edelberg, R. Electrodermal recovery rate, goal orientation, and aversion. *Psychophysiology*, 1972, 9, 512-520.
- Edelman, R. I. Effects of progressive relaxation on autonomic processes. *Journal of Clinical Psychology*, 1970, 26, 421-425.
- Freeman, G. L. The facilitative and inhibitory effects of muscular tension upon performance. *American Journal of Psychology*, 1933, 45, 17-52.
- Gellhorn, E. The physiological basis of neuromuscular relaxation. *Archives of Internal Medicine*, 1958, 102, 393-399.
- Graham, F. K., & Clifton, R. K. Heart rate change as a component of the orienting response. *Psychological Bulletin*, 1966, 65, 305-320.
- Haynes, S., Moseley, D., & McGowan, W. Relaxation training and biofeedback in the reduction of frontalis muscle tension. *Psychophysiology*, 1975, 12, 547-552.
- Haywood, H. C., & Spielberger, C. D. Palmar sweating as a function of individual differences in manifest anxiety. *Journal of Personality and Social Psychology*, 1966, 3, 103-105.
- Jacobson, E. *Progressive relaxation*. Chicago: University of Chicago Press, 1938.
- Jacobson, E. *Anxiety and tension control: A physiologic approach*. Philadelphia, Pa.: Lippincott, 1964.
- Jasper, H. H. Electrical signs of cortical activity. *Psychological Bulletin*, 1937, 34, 411-481.
- Katkin, E. S. The relationship between manifest anxiety and two indices of autonomic response to stress. *Journal of Personality and Social Psychology*, 1965, 2, 324-333.
- Katkin, E. S. The relationship between a measure of transitory anxiety and spontaneous autonomic activity. *Journal of Abnormal Psychology*, 1966, 71, 142-146.
- Kimble, G. A., & Posnick, G. M. Anxiety? *Journal of Personality and Social Psychology*, 1967, 7, 108-110.
- Lader, M. H., & Wing, L. *Physiological measures, sedative drugs, and morbid anxiety*. London: Oxford University Press, 1966.
- Lehrer, P. M. *The effect of progressive relaxation on habituation of the skin potential response: An analog of Wolpe's systematic desensitization*. Unpublished doctoral dissertation, Harvard University, 1969.
- Lehrer, P. M. Physiological effects of relaxation in a double-blind analog of desensitization. *Behavior Therapy*, 1972, 3, 193-208.
- Lehrer, P. M., & Taylor, H. A. Effects of alcohol on cardiac reactivity in alcoholics and nonalcoholics. *Quarterly Journal of Studies on Alcohol*, 1974, 35, 1044-1052.
- Lewinsohn, P. M. Some individual differences in physiological reactivity to stress. *Journal of Comparative and Physiological Psychology*, 1956, 40, 271-277.
- Lick, J. R., & Heffler, D. Relaxation training and tension placebo in the treatment of severe insomnia.

- nia. *Journal of Consulting and Clinical Psychology*, 1977, 45, 153-161.
- Maricq, H. R., & Edelberg, R. Electrodermal recovery rate in a schizophrenic population. *Psychophysiology*, 1975, 12, 630-633.
- McGuigan, F. J., Calvin, A. D., & Richardson, C. E. Manifest anxiety, palmar perspiration index, and stylus maze learning. *American Journal of Psychology*, 1959, 72, 434-438.
- McReynolds, P., Acker, M., & Brackbill, G. On the assessment of anxiety: IV. By measures of basal conductance and palmar sweat. *Psychological Reports*, 1966, 19, 347-356.
- Miller, M. Changes in the response to electric shock by varying muscular conditions. *Journal of Experimental Psychology*, 1926, 9, 26-44.
- Nicassio, P., & Bootzin, R. A comparison of progressive relaxation and autogenic training as treatments for insomnia. *Journal of Abnormal Psychology*, 1974, 83, 253-260.
- Nowlis, D., & Kamiya, J. The control of electroencephalographic alpha rhythms through auditory feedback and the associated mental activity. *Psychophysiology*, 1970, 6, 476-484.
- Orne, M. T., & Paskewitz, D. A. Aversive situational effects on alpha feedback training. *Science*, 1974, 186, 458-460.
- Paskewitz, D. A. A hybrid circuit to indicate the presence of alpha activity. *Psychophysiology*, 1971, 8, 107-112.
- Paskewitz, D. A., & Orne, M. T. Visual effects on alpha feedback training. *Science*, 1973, 181, 360-363.
- Paul, G. L. Inhibition of physiological response to stressful imagery by relaxation training and hypnotically suggested relaxation. *Behaviour Research and Therapy*, 1969, 7, 249-256. (a)
- Paul, G. L. Physiological effects of relaxation training and hypnotic suggestion. *Journal of Abnormal Psychology*, 1969, 74, 425-437. (b)
- Paul, G. L., & Trimble, R. W. Recorded vs. "live" relaxation training and hypnotic suggestion: Comparative effectiveness for reducing physiological activity and inhibiting stress response. *Behavior Therapy*, 1970, 1, 285-302.
- Rosenstein, A. J. Psychometric vs. physiological anxiety and serial learning. *Journal of Personality*, 1960, 28, 279-292.
- Routtenberg, A. The two-arousal hypothesis: Reticular formation and limbic system. *Psychological Review*, 1968, 75, 51-80.
- Shoemaker, J. S., & Tasto, D. L. The effects of muscle relaxation on blood pressure of essential hypertensives. *Behaviour Research and Therapy*, 1975, 13, 29-43.
- Sloane, R. B., Davidson, R. O., & Payne, R. W. Anxiety and arousal in psychoneurotic patients. *Archives of General Psychiatry*, 1965, 13, 19-23.
- Sokolov, E. N. *Perception and the conditioned reflex*. New York: Macmillan, 1963.
- Spelman, M. S. Psychological correlates of blood pressure. *Medical Journal of Australia*, 1966, 2, 1138-1140.
- Spence, K. W., & Taylor, J. A. Sex and anxiety differences in eyelid conditioning. *Psychological Bulletin*, 1966, 65, 137-142.
- Townsend, R. E., House, J. F., & Addario, D. A comparison of biofeedback-mediated relaxation and group therapy in the treatment of chronic anxiety. *American Journal of Psychiatry*, 1975, 132, 598-601.
- Wilson, A., & Wilson, A. S. Psychophysiological and learning correlates of anxiety and induced muscle relaxation. *Psychophysiology*, 1970, 6, 740-748.
- Woolfolk, R., Carr-Kaffashan, L., McNulty, T. F., & Lehrer, P. M. Meditation training as a treatment for insomnia. *Behavior Therapy*, 1976, 7, 359-365.

Received November 6, 1976 ■

Objective Measurement of Fear of Success and Fear of Failure: A Factor Analytic Approach

Susan Sadd, Michael Lenauer, Phillip Shaver, and Noel Dunivant
New York University

Several objectively scored measures of fear of success and fear of failure have been designed in recent years, but there is little evidence that they measure two distinct, unidimensional constructs. The present study was undertaken primarily to answer two questions: Are fear of success and fear of failure operationally distinct? Do all fear of success measures tap a single unidimensional construct? Eight fear of success and fear of failure scales were administered to 415 male and female subjects, and the scores were intercorrelated. Results indicated that fear of success is not a unidimensional construct and that some of the measures of fear of success and fear of failure are highly related. Next, each scale was factor analyzed, and 37 new variables were created. These were in turn factor analyzed, and five highly stable orthogonal factors were obtained. One of these factors appears to be fear of success; another is clearly test anxiety (called fear of failure in the literature on achievement motivation). A third factor is concerned with sex-role-related attitudes toward success in medical school. A fourth seems to reflect neurotic insecurity, and the fifth has to do with the value of success. Indices of psychological well-being and psychosomatic illness related differently to each of the five factors. Implications and further questions are discussed briefly.

Horner's research on fear of success (Horner, 1969a, 1969b, 1972, 1974) has generated well over 100 research articles in the past few years. According to Horner (1969a), fear of success (hereafter FOS) is "the fear that success in competitive situations will lead to negative consequences" (p. 38). Taken by itself this definitional phrase is quite general. However, because Horner was primarily interested in a form of FOS that she thought was characteristic of females in American

society, she included in her definition not only general negative consequences of success, such as social rejection ("unpopularity"), but also "loss of femininity." Most subsequent studies have failed to support this specific definition, however. In a recent review of the burgeoning literature on fear of success, Tresemer (1976) concluded that the "hypothesis that there is a gender difference in FOS is not supported" (p. 217). Therefore, the more general definition of FOS, leaving out loss of femininity, appears to be the most appropriate starting point for anyone undertaking studies in this area. Unfortunately, interested researchers will find little consensus concerning the best way to measure FOS. The present article takes a few steps toward remedying that problem.

Horner's earliest measure of FOS, and the one used most often by subsequent researchers, was a Thematic Apperception Test (TAT) type of projective device based on the following story line: "After first term finals

The research reported here was supported in part by a grant to Phillip Shaver from the Spencer Foundation, administered through Columbia University's Teachers College, and in part by a grant to Phillip Shaver and Katherine Garner from New York University. The authors are grateful to Katherine Garner and Caryl Dobrer for help in early phases of the study.

Requests for reprints should be sent to Phillip Shaver, Department of Psychology, New York University, 6 Washington Place, 7th floor, New York, New York 10003.

Anne (John) finds herself (himself) at the top of her (his) medical school class." Subjects were asked to complete the story by describing Anne or John in more detail, including how she (he) felt about this situation and what would probably happen in the future. This measure is problematic for several reasons. It is highly specific to a set of circumstances—success in medical school—which is obviously not synonymous with success in general. Moreover, until recently, American medical schools were masculine institutions, which meant that the hypothetical Anne was not only successful in a general sense but also highly deviant for her gender in terms of career choice (Lockheed, 1975). Add to this the fact that males were given the John cue and females the Anne cue, and it is easy to see that fairly specific attitudes about a successful male or female *in medical school* were being tapped, not necessarily a general personality trait, FOS. Finally, a present/absent scoring system based on one such story is not likely to be reliable (cf. Nunnally, 1967).

Because of these and other problems, many new measures of FOS have been developed recently, one by Horner and her colleagues (Horner, Tresemer, Berens, & Watson, Note 1) and several by other investigators (e.g., Cohen, 1975; Good & Good, 1973; Pappo, 1973; Spence, 1974; Zuckerman & Allison, 1976). Shaver (1976) reviewed the measures briefly and raised several questions about them. First, are all of the instruments measuring the same thing? The answer is almost certainly no. With some of the measures (e.g., Good & Good, 1973; Spence, 1974; Zuckerman & Allison, 1976), significant sex differences have been obtained, but this is not the case with other measures (Cohen, 1975; Pappo, 1973). Even more disturbing, the sex difference between means is not always in the same direction for different measures; Spence's instrument reveals more hostility among males toward a successful male, for example. Also disturbing are the low correlations between some of the new measures (e.g., Zuckerman & Allison, 1976) and Horner's original one, even though these are purported to measure the same construct.

A less familiar but even more serious problem is that FOS may not be different from what achievement motivation researchers (e.g., Atkinson & Feather, 1966) have called fear of failure. Shaver (1976) argued that most of the experimental results obtained by Cohen (1975), Pappo (1973), and Zuckerman and Allison (1976), which these authors interpreted as evidence for the construct validity of their FOS measure, could have been explained just as easily in terms of fear of failure. Compatible with Shaver's argument is the highly significant correlation of .57, which Pappo reported between her measure and the Debilitating Anxiety scale of the Achievement Anxiety Test (AAT; Alpert & Haber, 1960), a commonly used measure of fear of failure. Jackaway and Teevan (1976) compared the new FOS measure developed by Horner et al. (Note 1) with the projective measure of fear of failure devised earlier by Birney, Burdick, and Teevan (1969) and found significant correlations between the two measures for both males and females (.42 and .57, respectively). This finding is not surprising given the considerable overlap between the two scoring systems.

Believing that much of the inconsistency and confusion in the recent FOS literature is due to measurement problems, we undertook a factor analytic study of several measures of FOS and fear of failure. The study was designed primarily to answer two questions: (a) Are FOS and fear of failure measurably distinct? (b) To what extent do the recently devised objective tests of FOS measure a unitary (i.e., unidimensional) construct? To begin exploring another issue raised by Shaver (1976)—namely, that avoidance motives proposed in the achievement literature might be related to psychosomatic and psychological symptoms of conflict and stress, in addition to or instead of, being related to performance decrements—we also included questions concerning these symptoms.

The present article focuses only on objectively scored measures of FOS and fear of failure, because these are likely to be more reliable and are obviously easier to administer and score. Further research will be necessary to discover whether one or more of the di-

mensions we discovered are present in the projective measures as well. We did, however, include one measure that is directly related to Horner's original TAT-like device, Spence's (1974) objectively scored version of the Anne/John cue.

Measures Included in the Factor Analysis

Before presenting our study in detail, each measure will be described briefly. Five measures of FOS and two measures of fear of failure were included.

Spence (1974). The objective measure of FOS most similar to Horner's original projective measure is the one developed by Spence, which requires that subjects complete three stories, including the one about Anne (or John) in medical school, and then answer a series of multiple-choice questions about the story. For the Anne story these questions include: "How likable do Anne's classmates consider her?" In a study of 328 male and female undergraduates, Spence (1974) found a significant relationship between her measure and Horner's. Other analyses led Spence to conclude that her measure, as well as Horner's, tapped attitudes toward certain kinds of achievement rather than a deeply rooted personality trait. It is also worth mentioning, because we replicated Spence's results, that she found males to be more negative toward a successful male than females were toward a successful female. Although this is quite different from what Horner (1969b) would have predicted, it is compatible with many studies based on Horner's measure (Tresemer, 1974, 1976).

In the research to be reported here, we used the Anne (or John) story lead followed by 15 of Spence's multiple-choice questions.

Good and Good (1973). These authors developed a 29-item self-report measure of FOS based on Horner's (1969a, 1969b) notion that a person who fears success will be anxious and worried about other people's negative reactions to her (or his) success. In a sample of 228 male and female college undergraduates, the mean FOS score for females was significantly higher than the mean score for males, as Horner expected. The measure contains items of the following sort, each of

which is answered on a true-false scale: "I worry that I may become so knowledgeable that others will not like me." "If I were to do well at something, I would worry that someone might try to undermine my success."

Zuckerman and Allison (1976). These authors' 27-item scale contains statements concerning (a) benefits of success, (b) costs of success, and (c) attitudes toward success as compared with other alternatives. It was constructed on the basis of Horner's theorizing. The following items, answered on a 7-point agree-disagree scale, are representative: "The cost of success is overwhelming responsibility." "When competing against another person, I sometimes feel better if I lose than if I win." Based on Horner's description of FOS, Zuckerman and Allison predicted that FOS would be greater for females than for males, and in three different samples of undergraduate college students, females scored higher than males. (The difference was statistically significant in two out of three samples.) In two of their samples, Zuckerman and Allison's scale correlated weakly but significantly with Horner's (1969b) measure. In a validation study, high-FOS subjects performed significantly less well on an anagrams task than did low-FOS subjects.

Pappo (1973). The measures developed by Pappo (1973) and Cohen (1975) are different in conception from the ones described thus far. Whereas Horner, Spence, Good and Good, and Zuckerman and Allison were all interested in a motivational construct related to sex role socialization, Pappo and Cohen were interested in a neurotic form of FOS that might be equally prevalent among males and females. The two researchers differed slightly in their conceptualization of the etiology of neurotic FOS, with Pappo favoring Sullivanian theory (Sullivan, 1953) and Cohen basing her conception on Freud's discussion of oedipal conflicts.

Pappo developed an 83-item true-false questionnaire to measure "academic fear of success." The scale has been administered to large samples of high school and college students, and as expected no sex differences have been found. In an experimental study involving college students (Pappo, 1973), high

scorers on the scale exhibited poor performance on a reading test following success feedback; low scorers, in contrast, improved. The following are typical items from Pappo's scale: "I feel I need someone to push me to do the things I want to do." "I find it difficult to measure up to the standards I set for myself."

Cohen (1975). Cohen used Pappo's (1973) scale as a model; consequently, many of the items in the two scales are quite similar, although Cohen's are more general (i.e., not tied to academic situations). In a high-ability, achievement-oriented sample of high school students, Cohen's 64-item true-false scale successfully predicted which subjects would perform poorly on a memory task following success feedback. (Both Pappo's and Cohen's scales can be found in Canavan-Gumpert, Garner, & Gumpert, 1977.)

Measures of fear of failure. We included two measures of test anxiety or fear of failure, the Test Anxiety Scale (TAS) used in many studies by Sarason (e.g., 1972, which includes the scale items) and the AAT, designed by Alpert and Haber (1960) within the framework of achievement motivation theory (Atkinson & Feather, 1966). The TAS is a 37-item true-false scale. The AAT contains two subscales, one measuring debilitating anxiety (10 items, hereafter referred to as AAT-) and the other measuring facilitating anxiety (9 items, AAT+); both refer explicitly to academic or intellectual testing situations. The following items are typical of the TAS and the AAT-: "While taking an important exam I find myself thinking how much brighter the other students are than I am." "Nervousness while taking an exam or test hinders me from doing well." (The AAT+ is not important to later conclusions; essentially, it measures a form of arousal in test situations that is negatively correlated with debilitating test anxiety.)

Overview of the Present Study

We administered all of the measures described above to 415 university students; in addition, each subject provided background information and answered questions concerning psychological problems and psychosomatic

symptoms. Analysis of the data proceeded in stages. First, the reliabilities of the scales were computed for our sample, and, since these were acceptable, mean scores for males and females were compared. Second, the scales were intercorrelated, separately for males and females, to determine how strongly they related to each other. Third, each scale was factor analyzed, and 37 subscales were created. These subscales were in turn factor analyzed, yielding five orthogonal factors. Finally, scores on these five factors were correlated with background variables and measures of psychological well-being.

Method

Subjects

A letter was sent in the spring of 1976 to juniors and seniors at New York University inviting them to participate in a study of "achievement-related motivation and career choice." Older students were preferred, because they were closer to graduation and career choice, and thus to real-world success or failure. To obtain a sufficiently large sample, subjects were also recruited in classrooms and dormitory lobbies. Each subject received \$3 for filling out the battery of questionnaires. Data were collected from 430 subjects, but 15 of these failed to follow instructions or left out large numbers of items. Thus, 166 males and 249 females ($N = 415$) were represented in the data analyses.

Materials and Procedure

Each subject received a large envelope containing a background information form and two booklets of questions (Parts 1 and 2). Included in the background information form were standard demographic questions, questions about the subject's relationship with his or her parents, and a list of psychological problems and psychosomatic symptoms. Part 1 contained Horner's (1969b) medical school cue, followed by Spence's (1974) objective questions. (Females wrote about Anne in medical school; males, about John.) Instructions were the same as Horner's, except for a request that subjects were not to look at the multiple-choice questions following the story until after the story had been completed. Part 2 contained questions from the objective measures of FOS and fear of failure (Alpert & Haber, 1960; Cohen, 1975; Good & Good, 1973; Pappo, 1973; Sarason, 1972; Zuckerman & Allison, 1976). The items from all six measures were shuffled and printed in random order, and all were answered on a 4-point scale from strongly disagree, disagree, agree, to strongly agree. For items worded in the FOS or fear of failure direction, strongly disagree was

scored as 1; disagree, as 2; agree, as 3; and strongly agree, as 4. For items worded in the opposite direction (to control for acquiescence response bias), scoring was reversed. Scale scores were obtained by summing across items.

Data Analysis

Because of the complicated nature of the analyses to be reported, each stage of the analysis and its results will be described and discussed in turn. General discussion will be reversed until all analyses have been presented.

Results

Phase 1: Analyses Involving the Original Scales

Because scoring of the scales was modified slightly in the present study, so that all questions could be answered on the same 4-point continuum, internal-consistency reliability coefficients were computed (see Table 1). Since the number of items differs from scale to scale, estimated reliabilities for tests of standard length (20 items) were computed using results from the present study. This allows for comparison of the reliabilities of the various scales. These estimated reliabilities are also presented in Table 1.

Some of the scale constructors, following Horner (1969b), predicted sex differences in FOS (Good & Good, 1973; Spence, 1974; Zuckerman & Allison, 1976); others did not (Cohen, 1975; Pappo, 1973). In the present study, the only measure that produced a significant difference between males and females was Spence's measure of FOS. The results of the Spence questionnaire were similar to those reported by her in 1974; males showed significantly higher FOS than females. (Standard scores were computed for each of the 15 items on the Spence scale. These standard scores were summed to yield a total scale score for each subject; scores ranged from -15.17 to 28.58.) Means, standard deviations, and *t* tests are summarized in Table 2.

Of major interest in this study were the relationships among the various instruments used to measure FOS and fear of failure. A matrix of correlations among the eight scales was obtained separately for males and females. As can be seen by comparing the val-

Table 1
Reliabilities of the Original Scales for the Present Sample

Item	Coefficient α	Reliabilities of a 20-item scale
Fear of success		
Spence (1974)	.83	.87
Good & Good (1973)	.89	.84
Zuckerman & Allison (1976)	.68	.61
Pappo (1973)	.90	.69
Cohen (1975)	.92	.78
Fear of failure		
Debilitating scale	.84	.91
Facilitating scale	.64	.80
Test Anxiety Scale	.91	.85

ues for males and females in Table 3, there are very few differences between the corresponding correlations for the two sexes. A number of interesting relationships appeared for both males and females. For ease of understanding, these will be discussed under three separate headings: (a) relationships among two or more FOS measures; (b) relationships among the fear of failure measures; and (c) relationships between FOS and fear of failure. Each group of relationships will be discussed in turn.

1. Spence's (1974) measure of FOS correlated most strongly with the measure developed by Zuckerman and Allison (1976). ($r_t = .41$; $r_m = .37$; the subscripts *f* and *m* refer to males and females.) This finding is quite reasonable given that both Spence and Zuckerman and Allison designed their measures with Horner's work in mind. Nevertheless, the two measures are quite different in design and item format, so it is not surprising that the correlations between the two scales were not higher. It is impossible to determine the extent to which method variance, as distinct from differences in content, contributed to lowering these correlations.

The correlation between Pappo's (1973) and Cohen's (1975) measures approached the reliabilities of the scales ($r_t = .88$; $r_m = .86$). The correlations between Pappo's measure and other variables were quite similar to the

Table 2

Mean Scores and Standard Deviations for Males and Females on the Original Scales

Item	Males		Females		<i>t</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Fear of success					
Spence (1974)	2.47	8.12	-2.15	7.58	5.71*
Good & Good (1973)	61.28	10.53	59.83	11.08	1.29
Zuckerman & Allison (1976)	64.57	7.06	64.65	6.27	0
Pappo (1973)	197.40	20.72	196.74	22.81	.09
Cohen (1975)	157.90	18.83	157.80	20.39	0
Fear of failure					
Debilitating scale	22.93	4.72	23.53	4.97	1.20
Facilitating scale	21.95	3.37	22.03	3.39	.22
Test Anxiety Scale	87.85	13.14	90.02	15.17	1.47

* $p < .001$; all other *t* tests were nonsignificant.

correlations between Cohen's measure and those same variables. According to Pappo and Cohen, their conception of FOS is quite different from Horner's, a claim that is supported by the relatively low correlations between the Pappo and Cohen measures on the one hand and the measures developed from Horner's theory (Spence, 1974; Zuckerman & Allison, 1976) on the other. For example, Pappo's scale correlated .39 for females (.41 for males) with Zuckerman and Allison's scale: $r_t = .30$ with the measure designed by Spence ($r_m = .27$).

Good and Good's (1973) scale appears to measure elements common to the Horner-based scales and the measures designed by Pappo and Cohen. The correlations between FOS as measured by Good and Good and either the Horner-based measures or the Pappo and Cohen measures were higher than

any correlation between a Horner-based measure and Pappo's or Cohen's scale. Apparently, the Good and Good measure shares some content with both kinds of FOS measures.

2. The two measures of fear of failure, TAS and AAT—, were highly correlated with each other ($r_t = .83$, $r_m = .78$), thus replicating the findings of earlier investigators (Alpert & Haber, 1960; Sarason, 1960). The relationships between the TAS and other measures were nearly identical to the correlations between the AAT— and those same measures, which lends support to the contention that the TAS and AAT— measure the same construct. The AAT+ correlated negatively with the TAS and AAT— at about the level reported by Alpert and Haber (1960).

3. As mentioned earlier, some authors have suggested that FOS and fear of failure are indistinguishable motives (Jackaway & Tee-

Table 3

Correlations Among Eight Original Scales

Item	1	2	3	4	5	6	7	8
1. Spence (1974)	—	.36	.41	.30	.26	.24	.06	.25
2. Good & Good (1973)	.28	—	.67	.64	.64	.54	-.06	.50
3. Zuckerman & Allison (1976)	.37	.48	—	.40	.39	.29	-.06	.28
4. Pappo (1973)	.27	.69	.40	—	.88	.70	-.16	.69
5. Cohen (1975)	.27	.67	.48	.86	—	.66	-.24	.68
6. Debilitating scale	.19	.53	.20	.66	.62	—	-.40	.83
7. Facilitating scale	-.05	-.16	-.22	-.33	-.34	-.45	—	-.41
8. Test Anxiety Scale	.08	.50	.17	.62	.58	.78	-.46	—

Note. Correlations above the diagonal are for females ($n = 249$); those below the diagonal are for males ($n = 166$). For $n = 249$, $r_{.05} = .16$; for $n = 166$, $r_{.05} = .19$.

van, 1976; Shaver, 1976). The Pappo (1973) and Cohen (1975) measures correlated quite strongly with the measures of fear of failure, but the Horner-based measures did not. Since the latter are based on the notion that FOS develops through sex role socialization, one would not expect the Horner-based measures of FOS to be related to fear of failure. Nor would one expect them to be related if the Horner-Spence measure is tapping attitudes about success in medical school rather than fear of success.

Our examination of the correlations among the scales suggests that FOS is not unidimensional and that there is considerable commonality between FOS and fear of failure as they are currently operationalized. A factor analytic study is needed to reveal the basic underlying dimensions that have until now been confused.

Phase 2: Factor Analyses

To conduct a reliable factor analysis of 254 items, about 1,300 subjects would be needed, and this number was beyond our resources. Therefore, an alternative two-step procedure was followed: First, each of the eight scales (treating AAT+ and AAT- as separate scales) were factor analyzed; next, the resulting factor scores were factor analyzed. Since only 37 factors resulted from the first step of this procedure, it was possible to produce very stable factors in the second step using data from 415 subjects.

First-order factor analysis. Each of the eight scales was factor analyzed using a principal components analysis with communalities in the diagonal. (This and all other procedures followed in the study have been described by Gorsuch, 1974.) The number of factors extracted from each scale was determined using multiple criteria: the scree test, percent of variance accounted for, and interpretability. The extracted factors were rotated to the direct oblimin criterion.¹ An item with a factor loading greater than .30 was selected as an item representing that factor. When an item loaded significantly on two or more factors, the differences between loadings were tested for significance. If the difference was not significant, the item was assigned to

both factors in question. In the few cases in which items did not load significantly on any factor, they were eliminated from later analyses.

The first-order factor analysis yielded 9 factors for Pappo's (1973) scale, 4 for Zuckerman and Allison's (1976), 4 for Good and Good's (1973), 8 for Cohen's (1975), 6 for the TAS, 1 for AAT-, 1 for AAT+, and 4 for Spence's measure; thus, a total of 37 new variables were created.

Second-order factor analysis. Scores for each subject were computed on the 37 factors identified in the first-order factor analyses. All items that loaded .30 and above were assigned unit weights and were summed to produce total scores. To evaluate the stability of the factor solution, subjects were randomly assigned to two separate subsamples (Data Sets 1 and 2, or DS-1 and DS-2). Since there were significant mean differences between male and female scores on 2 of the 37 factors (1 from Spence's measure and 1 from Pappo's), within-groups correlations were computed for the two subsamples as recommended by Cooley and Lohnes (1971). Following an initial principal components analysis of the within-groups correlation matrix for DS-1, which produced a six-factor solution, the analysis was repeated specifying that it produce three-, four-, and five-factor solutions, each followed by both oblique and orthogonal (varimax) rotations. The choice of the final solution, 5 orthogonal factors, was based on clarity and interpretability.

To determine the degree of stability of the factor structure from DS-1, we attempted to replicate it using data from the second half of the sample (DS-2). A principal components analysis of the within-groups correlations for DS-2 was performed. Then, taking the varimax rotated factor structure matrix from DS-1 as the target matrix, the initial factor solution for DS-2 was subjected to an orthogonal Procrustes rotation (Cliff, 1966).²

¹ Factor analyses were computed using program FACTOR from the *Statistical Package for the Social Sciences* (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).

² These analyses were conducted using VARAN 2, a linear model variance analysis program (Hall, Kornhauser, & Thayer, 1973).

That this procedure succeeded in generating highly congruent factor structures is evinced in the correlations between matching factors for the two data sets: .97, .98, .92, .85, and .84. Having replicated the five-factor solution, we recombined the data from the two halves of the sample and computed a final five-factor varimax solution using all 415 cases. Of course, the results of this analysis were virtually identical to the results for each half of the sample.

*The new scales.*³ The end result of the factor analytic process can be viewed as five new scales composed of items from the original measures of FOS and fear of failure. To understand the scales and their implications, one must know something about their composition and meaning.

Factor 1 is composed of first-order factors from the measures designed by Good and Good (1973), Zuckerman and Allison (1976), Pappo (1973), and Cohen (1975). We have named this factor *Concern about the Negative Consequences of Success*, because all of the items loading highly on it reflect concern about jealousy, exploitation, criticism, sabotage, rejection, burdensome responsibility, and pressure following success. Also represented on this factor are statements about feelings and behaviors that may result from such concern. The factor is characterized by the following sorts of items:

(a) If I were outstanding at something, I would worry about the possibility of others making fun of me behind my back, and (b) the cost of success is overwhelming responsibility.

Of the five factors, this one comes closest to the general definition of FOS offered by Horner (1969a): "fear that success in competitive situations will lead to negative consequences" (p. 38).

Factor 2 is composed solely of first-order factors from Pappo's (1973) and Cohen's (1975) measures. This factor has been labeled *Self-deprecation and Insecurity*, and it is characterized by failure to live up to one's own standards, self-consciousness, unassertiveness, and behavioral manifestations of insecurity. This factor includes items such as: (a) I frequently find it difficult to measure up to the standards I set for myself, and (b)

I often brood about something I've said which may have been taken the wrong way by another person. It is extremely important that none of the items on this factor explicitly mentions negative consequences of success, and therefore the factor may not represent fear of success at all.

Factor 3 is simply *Test Anxiety*. All six first-order TAS factors, plus AAT- and AAT+, load on this factor. None of the components of the other scales loads significantly on Factor 3. This is probably due in part to the extreme specificity of the situation described in the scales; all refer directly to anxiety in testing situations. Whether this factor deserves the more general label of fear of failure is disputable.

Factor 4 contains all of the first-order factors derived from Spence's measure of FOS, with no contributions from any other measure. This too is probably the result of the specificity of the situation described in the measure, which leads us to call it *Attitudes Toward Success in Medical School*. This designation is compatible with Spence's description of Horner's cue as a measure of attitudes rather than personality.

Factor 5 is called *Extrinsic Motivation to Excel*. Items on this factor are concerned with the extreme importance of success, status, and power. There is no mention of fear or negativity of success. The 19 items come from two factors of Zuckerman and Allison's (1976) scale and one factor from Pappo's (1973) scale; they include (a) When you're on top, everyone looks up to you, and (b) I feel that it is important for people of higher status to like me. Most of the items on this factor were written to reflect the absence or opposite of FOS; in fact, however, they seem to measure an independent dimension.

Phase 3: Relationships Between the Factors and Psychological Well-Being and Psychosomatic Symptoms

Five factor scores were computed for each subject using the varimax factor weights ob-

³ A complete list of items, factor loadings, and scoring procedures can be obtained from the authors on request.

tained in the factor analysis based on all 415 subjects. These scores were then correlated with demographic variables, family background variables, grade point average, and indices of psychological well-being including psychosomatic symptoms.⁴ The latter indices were based on a 4-point scale associated with the following question: "How much have the following problems bothered you during the past year?" (The alternatives were not at all, a little bit, moderately, and quite a bit.) We expected the first three factors—Concern over Negative Consequences of Success, Self-deprecation and Insecurity, and Test Anxiety—to be related to psychological and psychosomatic symptoms, since all of them involve anxiety or conflict. This should be especially true for the Self-deprecation factor, which is very similar to measures of general anxiety or neurosis. We did not expect Factor 4, Attitude Toward Success in Medical School, to be related to psychological well-being, although Horner's (1969b, 1974) original work might have led some investigators to expect this. Factor 5, Extrinsic Motivation to Excel, might be related to a few symptoms of stress, since it connotes great drive and ambition, but there is no certain basis for this prediction. (Factor 5 was unanticipated and is still the least well understood of the five factors.)

The criterion for considering a correlation coefficient as meaningful was set at .20. Although small, a correlation of this size is highly reliable ($p < .001$). All correlations above .20 are shown in Table 4.

As expected, Factor 4 (Attitudes Toward Success in Medical School) was not correlated with any of the psychological or psychosomatic variables. Factor 5, Extrinsic Motivation to Excel, was correlated with only two items, both indicating sleep disturbances. Also as expected, Factor 2, Self-deprecation and Insecurity, was related to the largest number of problems and symptoms (14) and yielded the largest correlation coefficients. Notice that the highest correlation in this list is the one between Factor 2 and feelings of worthlessness, which is quite consistent with the label assigned to the factor Self-deprecation and Insecurity. Test anxiety (Factor 3) was

Table 4
Correlations of Factors with Psychosomatic Symptoms

Symptom	Factor			
	1	2	3	5
Headaches				.20
Feeling tired or low in energy		.36		
Poor appetite	.20			
Crying easily		.20	.21	
Feeling lonely		.39		
Worry and anxiety		.39	.23	
Irrational fears		.37	.21	
Nausea or upset stomach		.22		
Trouble falling asleep			.21	.22
Sleep that is restless or disturbed		.28	.28	.20
Waking up too early in the morning	.24			
Feeling tense or keyed up		.37		
Overeating		.25		
Feelings of worthlessness		.53	.29	
Feelings of guilt		.40	.24	
Trouble concentrating		.40	.34	
Feeling that you just can't go on	.26	.31	.30	
Difficulty making decisions		.43		

Note. None of the correlations with Factor 4 exceeds .20. The following symptoms did not correlate above .20 with any factor: faintness or dizziness; loss of sexual interest or pleasure; stomach ulcers or colitis; pains in heart or chest; pains in lower back; recurring diarrhea; chronic constipation; high blood pressure; trouble getting your breath.

also correlated with several psychosomatic and psychological problems, although not as many as were related to Factor 2. The fact that the highest relationship was between test anxiety and poor concentration is quite compatible with previous research on test anxiety (e.g., Sarason, 1972; Wine, 1971). Factor 1, Concern over the Negative Consequences of Success, which we have said seems closest to Horner's original conception of FOS, was also related to three psychological and psychosomatic problems: poor appetite, waking up too early in the morning, and feeling that one "just can't go on." Perhaps it is worth mentioning that all three are common symptoms of depression.

⁴ Correlations with demographic and background variables will not be discussed in this article. A list and discussion of the significant relationships can be obtained from the authors.

Discussion

We began with two major questions: (a) Are all of the recently proposed objective tests of FOS measuring the same unidimensional construct? and (b) Are FOS and fear of failure measurably distinct? The answer to the first question is no. The answer to the second question is more complex. FOS as measured by Pappo (1973) and Cohen (1975) is closely related to measures of test anxiety (fear of failure), but FOS as measured by Spence is not.

Seeking clearer answers to our two major questions, we factor analyzed each of the original scales, producing 37 new variables. These variables were then factor analyzed, and five orthogonal factors were produced. One of these, Factor 3, was clearly Test Anxiety; all components of the original test anxiety scales (TAS, ATT+, and ATT-) loaded on this factor, but no other variables did. If one agrees with achievement motivation researchers that test anxiety should be called "fear of failure," then Factor 3 represents fear of failure. We are hesitant to accept this general label, however. The items on Factor 3 are quite specific to academic or intellectual tests and may not be good indicators of fear of failure in other situations.

Factor 2, which we have called Self-deprecation and Insecurity, might serve as a more general measure of fear of failure. The items reflect fear that the person cannot or will not live up to his or her own standards. Some of the items refer explicitly to low self-confidence and anticipation of failure. As defined by achievement-motivation researchers, fear of failure (or the "motive to avoid failure") is a "capacity to experience shame given non-attainment of a goal (failure)" (Weiner, 1972, p. 200). High scorers on Factor 2 surely display this capacity.

There are two obstacles to accepting Factor 2 as a general measure of fear of failure, however. First, many of the items that load highly on Factor 2 deal with neurotic inhibition of assertiveness; high scorers are afraid to express their desires or to stand out in a group. We do not know for sure whether this should be included in a general measure of fear of failure. Second, Pappo (1973) and

Cohen (1975) argued that FOS is unconscious and hence is expressed only indirectly; it may even manifest itself to the success-anxious person as fear of failure. This argument is weakened somewhat by our finding that the items that explicitly mention negative reactions to success load on Factor 1, not on Factor 2, which contains the largest portions of Pappo's and Cohen's scales. Still, we cannot completely rule out the possibility that Factor 2 measures unconscious fear of success. Only careful experimental studies can clarify this matter.

Factor 1, Concern over Negative Consequences of Success, fits very well with Horner's original conception of FOS, except that it does not include loss of femininity. Horner believed that FOS was significantly more prevalent in women than in men because it was the result of sex role socialization processes. According to the stereotypic image of males and females, independence, competence, intellectual achievement, and leadership are positive attributes for males, but they are inconsistent with femininity. In the present study, however, we found no sex differences in concern about negative consequences of success. Males are just as likely as females to exhibit concern over the negative consequences of success. None of the items that load on Factor 1 mentions conflicts with or loss of femininity. Rather, the concern is over jealousy, exploitation, social rejection, and excessive pressure and responsibility. These are as likely to be concerns of men as they are of women. Thus, contrary to Horner's (1972, 1974) theory, sex role socialization is not sufficient to explain the development of this form of FOS. The real developmental factors that cause some people, male or female, to score high on Factor 1 remain to be identified.

We should not ignore the possibility that Horner was correct about the link between femininity and FOS in 1968. (Her data were collected in 1965.) The social climate has changed considerably since then, and this certainly could have affected our subjects, who were college juniors and seniors in 1976. In fact, Tresemer (1976) presented evidence for a historical change. Despite changes in sex role orientation, and perhaps also in atti-

tudes toward achievement, however, we still find many males and females who experience FOS as measured by Factor 1.

The factor closest to Horner's (1969b) original measure of FOS is Factor 4, which we have called Attitudes Toward Success in Medical School to indicate that we doubt whether it has much to do with fear of success. In our study, as in Spence's (1974), males were actually more negative toward a successful male medical student than females were toward a corresponding female. Spence showed that the reasons for the hostility were not identical for males and females. Thus, some of Horner's arguments might be correct despite what seems to be a strongly contradictory pattern of results. However, given that the attitudes expressed in people's stories are quite complex and probably specific to the medical school situation, it is not surprising that research based on this measure has been confusing and contradictory (Tresemer, 1976).

Factor 5, Extrinsic Motivation to Excel, was not anticipated and remains somewhat of a mystery. For the moment, the most important point to be made about Factor 5 is that it contains items from the scales designed by Pappo (1973) and Zuckerman and Allison (1976), who expected the items to indicate absence of FOS; in fact, they appear to measure an entirely different dimension.

Directions for further research are clear. (a) It would be useful to develop shorter scales to represent Factors 1, 2, and 5. Spence's measure and the existing test anxiety scales are acceptable in their original forms. (b) Experimental studies are needed to assess the behavioral consequences of FOS and fear of failure. FOS could be measured by the items on Factor 1; either Factor 2 or Factor 3 could be used to measure fear of failure, whichever proves to have the most construct validity. Presumably, a person for whom FOS is the dominant motive would perform more poorly following success than following failure. A person with strong fear of failure and little or no fear of success should be disorganized by failure but quite pleased by success. (c) The possibility that achievement conflicts lead to psychosomatic symptoms of specific kinds deserves much

more attention than it has received to date. (d) It is important to determine the developmental antecedents of the kind of FOS measured by Factor 1. (e) Since we did not include Horner's most recent measure (Horner et al., Note 1), it would be useful to conduct a study explicitly to determine how her new measure, which is beginning to be widely used, is related to the five dimensions discussed here. If Jackaway and Teevan (1976) are right, Horner's new measure will be substantially correlated with our fear of failure dimensions (Factors 2 and 3). (f) Finally, it is important to reevaluate Atkinson's model of achievement motivation (Atkinson & Feather, 1966) in light of FOS research. Horner intended for the concept to play a role in a revised comprehensive model of achievement behavior, but so far no one has proposed such a model.

Reference Note

1. Horner, M. S., Tresemer, D. D., Berens, A. E., & Watson, R. I., Jr. *Scoring manual for an empirically derived scoring system for motive to avoid success*. Unpublished manuscript, Harvard University, 1973.

References

- Alpert, R., & Haber, R. N. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, 1960, 61, 207-215.
- Atkinson, J. W., & Feather, N. T. (Eds.) *A theory of achievement motivation*. New York: Wiley, 1966.
- Birney, R., Burdick, H., & Teevan, R. *Fear of failure*. New York: Van Nostrand-Reinhold, 1969.
- Canavan-Gumpert, D., Garner, K., & Gumpert, P. *Fear of success: Theory and research*. Boston: Heath, 1977.
- Cliff, N. Orthogonal rotation to congruence. *Psychometrika*, 1966, 1, 33-42.
- Cohen, N. E. Explorations in the fear of success. (Doctoral dissertation, Columbia University, 1974). *Dissertation Abstracts International*, 1975, 36, 1425B-1426B. (University Microfilms No. 75-18, 368)
- Cooley, W. W., & Lohnes, P. R. *Multivariate data analysis*. New York: Wiley, 1971.
- Good, L. R., & Good, K. C. An objective measure of the motive to avoid success. *Psychological Reports*, 1973, 33, 1009-1010.
- Gorsuch, R. L. *Factor analysis*. Philadelphia, Pa.: Saunders, 1974.
- Hall, C. E., Kornhauser, K., & Thayer, D. *VARAN 2: A linear model variance analysis program* (2nd

- ed., Research Memorandum RM-73-30). Princeton, N.J.: Educational Testing Service, 1973.
- Horner, M. S. Fail: Bright women. *Psychology Today*, 1969, 2, pp. 36-38; 62. (a)
- Horner, M. S. Sex differences in achievement motivation and performance in competitive and non-competitive situations. (Doctoral dissertation, University of Michigan, 1968). *Dissertation Abstracts International*, 1969, 30, 407B. (University Microfilms No. 69-12, 135) (b)
- Horner, M. S. Toward an understanding of achievement related conflicts in women. *Journal of Social Issues*, 1972, 28(2), 157-176.
- Horner, M. S. The measurement and behavioral implications of fear of success in women. In J. W. Atkinson & J. O. Raynor (Eds.), *Motivation and achievement*. New York: Wiley (Halsted Press), 1974.
- Jackaway, R., & Teevan, R. Fear of failure and fear of success: Two dimensions of the same motive. *Sex Roles*, 1976, 2, 283-293.
- Lockheed, M. E. Female motive to avoid success: A psychological barrier or a response to deviancy? *Sex Roles*, 1975, 1, 41-50.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. *Statistical package for the social sciences*. New York: McGraw-Hill, 1975.
- Nunnally, J. *Psychometric theory*. New York: McGraw-Hill, 1967.
- Pappo, M. Fear of success: A theoretical analysis and the construction and validation of a measuring instrument. (Doctoral dissertation, Columbia University, 1972). *Dissertation Abstracts International*, 1973, 34, 421B. (University Microfilms No. 73-16, 235)
- Sarason, I. G. Empirical findings and theoretical problems in the use of anxiety scales. *Psychological Bulletin*, 1960, 57, 403-415.
- Sarason, I. G. Experimental approaches to test anxiety: Attention and the uses of information. In C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research*. New York: Academic Press, 1972.
- Shaver, P. Questions concerning fear of success and its conceptual relatives. *Sex Roles*, 1976, 2, 305-320.
- Spence, J. T. The Thematic Apperception Test and attitudes toward achievement in women: A new look at the motive to avoid success and a new method of measurement. *Journal of Consulting and Clinical Psychology*, 1974, 42, 427-437.
- Sullivan, H. S. *The interpersonal theory of psychiatry*. New York: Norton, 1953.
- Tresemmer, D. Fear of success: Popular but unproven. *Psychology Today*, 1974, 7, 82-85.
- Tresemmer, D. The cumulative record of research on "fear of success." *Sex Roles*, 1976, 2, 217-236.
- Weiner, B. *Theories of motivation*. Chicago: Markham (Rand McNally), 1972.
- Wine, J. Test anxiety and direction of attention. *Psychological Bulletin*, 1971, 76, 92-104.
- Zuckerman, M., & Allison, S. N. An objective measure of fear of success: Construction and validation. *Journal of Personality Assessment*, 1976, 40, 422-430.

Received November 26, 1976 ■

Garfield Appointed Editor, 1979-1984

The Publications and Communications Board of the American Psychological Association announces the appointment of Sol L. Garfield as Editor of the *Journal of Consulting and Clinical Psychology* for the years 1979-1984. As of January 1, 1978, manuscripts should be directed to the Editor-elect:

Sol L. Garfield
Department of Psychology
Washington University
St. Louis, Missouri 63130

WISC-R Factor Structures Among Anglos, Blacks, Chicanos, and Native-American Papagos

Daniel J. Reschly
Iowa State University

Wechsler Intelligence Scale for Children-Revised (WISC-R) factor structures were compared for sample of Anglo, Black, Chicano, and Native-American Papago children from Pima County, Arizona. The samples were randomly selected from school enrollment rosters and stratified by ethnicity, grade level, sex, and urban-rural residence ($N = 950$). Application of two objective procedures for determining the appropriate number of factors for each group suggested a three-factor solution for Anglos, a two- or three-factor solution for Chicanos depending on procedure used, and two-factor solutions for Blacks and Native-American Papagos. The two-factor solutions were highly similar for the four groups. The three-factor solutions were similar for Anglos and Chicanos but were substantially different for the other groups. The groups were highly similar in terms of the proportion of variance accounted for by a general factor, and the Verbal-Performance scale distinction appeared equally appropriate for all groups.

Examination of the construct validity of a test in samples of diverse sociocultural groups provides evidence concerning the appropriateness and fairness of the use of the test with different groups. Comparability of factor analysis results for different groups, and the degree to which the results of the factor analysis are consistent with the major scores and common interpretations of the test are necessary conditions for fairness in use of the test with culturally diverse persons. Indeed, if a test is not measuring the same underlying abilities or if the commonly used scores from the test represent varying abilities depending on group membership, then use of the test with culturally different persons is probably

inappropriate and unfair, and the predictive validity of the test is likely to be lower for specific groups. Since the Wechsler Intelligence Scale for Children (WISC) and the Wechsler Intelligence Scale for Children-Revised (WISC-R) have been the most frequently used measures of general intelligence in schools and clinics, data from diverse groups concerning the construct validity of the WISC-R are needed for practitioners to make judgments concerning its appropriateness and possible fairness.

Although numerous investigations of the factor structure of the WISC appeared in the literature (Sattler, 1974), only three studies compared factor similarity among diverse sociocultural groups, and no studies on the WISC-R of this nature have appeared to date. Generally, these studies found high similarity for Blacks and Anglos (Lindsey, 1967) and for Blacks, Anglos, and Chicanos (Silverstein, 1973) when two-factor solutions for the WISC-R were used. However, when the three-factor solutions were applied (Semler & Iscoe, 1966), significant differences resulted in the second and third factors for Blacks and Anglos.

The collection of these data was supported by a grant from the Arizona State Department of Education, Division of Special Education, through Title VI-B, EHA (91-230) funds. The technical assistance provided by Larry Hanser in the analysis of the data and the cooperation and assistance of the school districts, state department personnel Karen Davis, Earl Pike, and Fred Jipson in the collection of the data are gratefully acknowledged.

Requests for reprints should be sent to Daniel J. Reschly, Department of Psychology, Iowa State University, Ames, Iowa 50011.

The present study attempted to extend Kaufman's (1975) factor analysis of the 1974 WISC-R to three non-Anglo groups. Kaufman interpreted three factors in an analysis of the standardization data by age level. These factors were labeled *Verbal Comprehension* (VC), formed by the subtests of Information (I), Similarities (S), Vocabulary (V), and Comprehension (C); *Perceptual Organization* (PO), formed by Picture Completion (PC), Picture Arrangement (PA), Block Design (BD), Object Assembly (OA), and Mazes (M); and *Freedom from Distractibility* (FD), formed by Arithmetic (A), Digit Span (DS), and Coding (Co). In addition, strong support from the factor analysis data was reported for Wechsler's use of the Full Scale IQ score as an index of general intelligence and for the organization of the test into the Verbal and Performance IQ Scales. Silverstein (1977) also factor analyzed the standardization data, and, although a slightly different method of rotating factors was used, the same three-factor pattern emerged. Thus, the WISC-R factor structure, in contrast to the WISC, appears to be highly stable across both age levels and different methods of conducting the analysis.

The purposes of the present study were to examine the appropriateness and fairness of the WISC-R for four sociocultural groups in terms of (a) comparability of factor structures; that is, does the WISC-R measure the same underlying abilities for Anglo and non-Anglo groups? and (b) construct validity evidence for the Full Scale IQ and the verbal-performance organization of the test; that is, does the Full Scale IQ measure general intelligence and do the Verbal and Performance IQ scores represent somewhat different but overlapping abilities for the four groups?

Method

Sample

In November 1973 the Division of Special Education, Arizona State Department of Education funded a comprehensive study of handicapping conditions among school-age children. The Pima County Special Services Cooperative was authorized to conduct the study in Pima County. The data reported herein were gathered as part of this study. Pima

County is geographically large (9,200 square miles), ethnically diverse (approximately 68% Anglo, 25% Chicano, 4% Black, and 3% Native American), and largely urban in population (Tucson) with extensive and sparsely populated rural areas.

A stratified random sample of 1,040 children was selected with equal numbers from each of the stratification variables of group (Anglo, Black, Chicano, and Native-American Papago, $n = 260$ per group), sex, urban-rural residence, and grade (first, third, fifth, seventh, and ninth; $n = 208$ per grade). The entire sample of Black children was urban, and the entire sample of Native-American Papago children was rural due to the very low proportions of urban Indians and rural Blacks in Pima County.

The cooperation of Tucson District 1, which enrolls about two thirds of all school-age children in the county, and all of the rural school districts in the county was obtained through contacts with district authorities. School district enrollment rosters were used to randomly select the sample. Ethnicity was determined by school data and in some cases by contacting the parents. Tucson District 1 was regarded as urban. Outlying districts, 25 miles or more from Tucson, were regarded as rural. Parents of children selected in the initial sample were contacted by letter or phone to explain the nature of the study and to solicit written permission. If parent permission was not obtained due to refusal (4%), no reply (18%), or the child withdrew from school, parents moved, and so forth (11%), another child was selected from an alternative sample constituted by the above process. There were no appreciable differences among the groups in percentages of parents granting permission, refusing permission, no reply, or no address—family moved, and so forth. Due to various logistical problems, for example, delays in return of parental permission, school scheduling problems, and availability of examiners to travel to remote areas, WISC-R scores were obtained for 950 of the original sample of 1,040 students (Anglo, $n = 252$; Black, $n = 235$; Mexican American, $n = 223$; and Native-American Papago, $n = 240$). The average age of the final sample was 10.63 years (range from 6.28 to 15.87 years) including 468 males and 482 females. Additional information concerning the sample is provided in Reschly and Jipson (1976).

As soon as parental permission was obtained, appointments were made with school officials to administer the various assessment procedures. The WISC-R was administered by appropriately trained examiners, and all WISC-R protocols were further checked by me for clerical and scoring errors.

Analysis

For each of the four ethnic groups, two procedures were used to provide a guide to the appropriate number of factors needed to efficiently and thoroughly describe the WISC-R. Following Silverstein's (1977) and Kaufman's (1975) methods, a principal components analysis was conducted with

is in diagonals. An eigenvalue greater than 1 was the criterion for determining the appropriate number of factors. Second, an unrestricted maximum likelihood analysis was conducted for the two-, three-, and four-factor solutions. At each step in the analysis, a chi-square goodness of fit test was conducted to compare the resulting factor matrix and the original correlational matrix.

A principal factor analysis with squared multiple correlations in the diagonals was then conducted separately for each group followed by varimax rotation of the two-, three-, and four-factor solutions. Varimax rotation of the two-, three-, and four-factor solutions was conducted regardless of the outcomes of the tests for appropriate number of factors to conform as closely as possible to the previous literature.

Results

Number of Factors

The techniques used as guides to the appropriate number of factors yielded somewhat inconsistent results. The principal components criterion (eigenvalues > 1) suggested three-factor solutions for Anglos and Chicanos and two-factor solutions for Blacks and Native-American Papagos. The chi-square tests for goodness of fit of the two-factor solutions, that is, testing the hypothesis that the resulting two-factor matrix accounts for the variation in the original matrix of simple correlations among subtests, suggested that more than two factors were required only for Anglos ($p < .07$) and that two factors were sufficient for the other groups ($p > .15$). Using Kaufman's (1975) reasoning, a three-factor solution might be explored for all of the groups, since there was evidence for the insufficiency of a two-factor solution for some of the groups, and the eigenvalues for the third factor in the three-factor solutions were very close to 1.0 for all groups. Although the ultimate criterion for the appropriate number of factors is psychological meaningfulness (including descriptive efficiency, statistical invariance, and predictive utility), the objective guides clearly suggested only two factors for Blacks and Native Americans and perhaps only two factors for Chicanos.

Comparison of Two-Factor Solutions

The two-factor solutions yielded the now familiar pattern of a first factor composed

of the Verbal scale subtests and a second factor constituted by the Performance scale subtests.¹ The subtests with the highest loadings on the first factor for all groups were Vocabulary (V), Information (I), Comprehension (C), and Similarities (S), although all the verbal subtests were significantly related to the first factor. (Loadings were .40 or above for all six subtests except for the loading of Digit Span (DS) for one group, $Mdn = .65$.) The second factor formed by Performance scale subtests was again highly similar across the four ethnic groups. The Block Design (BD) and Object Assembly subtests had the highest loadings for all groups. All Performance scale subtests were significantly related to the second factor (loadings of .40 or above) except for Coding (Co), which was not a significant component of the second factor for any of the groups (loading $< .30$ for all groups). Examination of coefficients of congruence between matching factors among the four groups further supported the judgment of very high similarity among the factors. The 12 coefficients of congruence varied from .97 to .99.

Comparison of Three-Factor Solutions

Three- and four-factor solutions were analyzed for all groups even though the objective criteria unequivocally supported three factors for Anglos only, either two or three factors for Chicanos depending on the criterion used, and only two factors for Blacks and Native-American Papagos. In addition to inspection of the size and pattern of the loadings on the factors, coefficients of congruence were computed among matching factors for the groups in this study and between the varimax median loadings reported for the standardization sample (Kaufman, 1975, Table 4, p. 141). The three-factor solution for Anglos (see Table 1) was virtually identical to the three-factor pattern reported for the standardization sample

¹ Tables of the subtest loadings for each group in the two-factor solutions and the coefficients of congruence for matching factors in the two- and three-factor solutions are available from the author. Also, the loadings of each subtest on the first unrotated factor are available in these tables.

Table 1

Wechsler Intelligence Scale for Children-Revised Subtest Loadings in Three-Factor Solution for Four Ethnic Groups

Subtest	Anglo			Black			Chicano			Native-American Papago		
	I	II	III	I	II	III	I	II	III	I	II	III
Information	63	32	26	66	40	18	66	20	33	68	22	21
Similarities	59	26	26	59	41	13	67	15	22	58	33	11
Arithmetic	43	26	45	61	34	27	40	13	45	42	37	09
Vocabulary	74	23	12	75	20	16	67	26	30	74	15	05
Comprehension	64	22	21	71	24	09	61	20	06	70	10	17
Digit Span	35	02	40	49	08	36	33	14	31	30	35	09
Picture Completion	20	49	09	25	52	21	32	52	12	21	53	14
Picture Arrangement	20	53	00	29	53	24	17	38	39	23	44	03
Block Design	17	60	22	20	33	58	20	59	16	14	69	05
Object Assembly	07	59	18	10	17	58	14	58	09	07	51	25
Coding	12	16	40	33	20	22	14	16	37	17	17	37
Mazes	18	42	10	23	44	30	06	47	20	14	51	28

Note. All decimal points have been omitted. Roman numerals refer to the factor number: Factor I = Verbal Comprehension; Factor II = Perceptual Organization; Factor III = Freedom from Distractibility.

by Kaufman (1975). The first factor was formed by the V, C, I, and S subtests. The A and DS subtests also had substantial loadings on the first factor, but their highest loadings were on the third factor. The second factor for Anglos was the familiar Perceptual Organization (PO) factor formed by the BD, OA, Picture Arrangement (PA), Picture Completion (PC), and Mazes (M) subtests. A third factor, previously described as Freedom From Distractibility (or memory, number, sequential, etc.) was formed by the A, DS, and Co subtests. Coefficients of congruence between the three-factor solutions for Anglos and the data reported by Kaufman were .98, .98, and .97, respectively, for the three factors, further supporting the conclusion of near-perfect replication of the three-factor patterns reported for the standardization sample.

The first two factors in the three-factor solution for Chicanos were highly similar to the above pattern, but the third factor was slightly different in that the loading on PA was slightly higher on the third rather than the second factor, and the loading of DS on the third factor was somewhat lower. Coefficients of congruence for the three factors, respectively, between the Chicano and Anglo loadings were .99, .98, and .86, and .98, .99, and .93 between the loadings for Chicanos and the standardization sample.

The three-factor solutions for Native-American Papagos and Blacks were clearly different from the previously reported three-factor solutions and should, perhaps, *not* be interpreted at all, since the objective guides to appropriate number of factors suggested the sufficiency of two-factor solutions. For both groups, the first two factors in the three-factor solutions were formed by the Verbal and Performance scale subtests, respectively. However, for Blacks the second factor was formed by PA, PC, and M and the third factor by BD and OA. Thus the third factor for Blacks appeared to involve a splitting of the second factor into two factors and clearly did not match the previously reported patterns for the third factor. A similar result was obtained for Native-American Papagos for whom the third factor was formed by one subtest only (Co), which suggests subtest-specific or method variance rather than a factor as such. However, it should be noted that the third factor for Native-American Papagos was similar to what Cohen (1959) called Factor E (Quasi-Specific), which also was formed primarily by Co. Coefficients of congruence for the three-factor solutions among all groups and the standardization data were very high for the first factor ($\geq .96$) and only slightly lower for the second factor ($\geq .89$). However, for the third factor, the coefficients for Blacks

and Native-American Papagos were significantly lower (.72-.78) in all comparisons, further supporting the conclusion of *not* interpreting a third factor for these groups.

The fourth factor in the four-factor solutions was uninterpretable for all groups. (All subtest loadings were $<.30$ with most near zero on the fourth factor.)

General Intelligence

The data were further analyzed to determine the degree to which a general factor of intelligence was measured by the WISC-R for the different groups. In addition to the obvious evidence of a general factor, that is, the significant and positive correlations of subtests with each other and with Full Scale score, three indices of a possible general factor were examined. First, following Kaufman's (1975) analysis, which resulted in a median of 82% of the common factor variance attributed to the general factor, the loadings on the unrotated first principal factor were analyzed separately for each group. The percentage of common factor variance accounted for by the general factor was nearly the same regardless of group (79, 83, 79, and 77, respectively, for Anglos, Blacks, Chicanos, and Native-American Papagos). A second analysis of the general factor using the first unrotated principal component resulted in about the same percentages for each group.

A third index, which I preferred, was based on a restricted maximum likelihood analysis that used a target matrix based on the results of the two-factor solutions for each group. This method provided a lower but more realistic estimate of the variance attributable to the general factor, because the unique variance associated with the Verbal Comprehension (VC) and PO factors (clearly the strongest factors for all groups) was taken from the general factor and apportioned to the respective VC and PO factors. This analysis yielded estimates of 61%, 63%, 59%, and 61% of the variance attributable to a general factor for Anglos, Blacks, Chicanos, and Native-American Papagos, respectively. Regardless of the index used, the proportions of variance attributable to a general factor were approxi-

mately the same for all groups and similar to the standardization sample.

Verbal-Performance Organization

The first and second factors, VC and PO, were highly similar to the respective Verbal and Performance scales in the two-factor solutions for all groups and in the three-factor solutions for all groups except Blacks. The closeness of the VC and PO factors to the verbal-performance dichotomy is even greater when the supplementary subtest for the verbal scale (DS) is ignored and when M is substituted for Co on the Performance scale. If these stipulations are followed, then the respective scales conform almost perfectly to the first and second factors for all groups.

Discussion

Some caution regarding the results of this study must be expressed, since in addition to race or ethnicity, the groups also varied significantly on socioeconomic status and level of intelligence. Due to limitations in sample size in this and most other investigations, it was impossible to analyze separately the possible effects of socioeconomic status, level of intelligence, and group membership. The well-known fact that these variables are not independent requires caution in interpreting the *differences* reported among the groups. A slight relationship between level of intelligence and factor patterns on the WISC and WISC-R has been reported previously (Van Hagen & Kaufman, 1975), and may partially account for the group differences reported here.

The major differences in factor patterns among the groups were essentially restricted to the question of the existence and composition of the third (Freedom from Distractibility—FD) factor. The number of interpretable factors on the WISC-R varied for these samples, and perhaps vary for these groups generally. The previously reported pattern of three factors for Anglos was clearly replicated, and the results for Chicanos were generally consistent with this pattern. However, the objective guides to the appropriate number of factors and the pattern of factor load-

ings in the three-factor solutions both failed to support the existence of the third (FD) factor for Blacks and Native-American Papagos. The existence of the third factor on the WISC was the subject of some debate (Silverstein, 1969), and the "meaning" of this factor has never been altogether clear. Some researchers (e.g., Cohen, 1957) regard the third factor as a measure of intellectual skills such as immediate memory, numerical or sequencing. Others, including Cohen (1959), have described the third factor in more behavioral or nonintellectual terms such as attention-concentration, or more commonly now, Freedom from Distractibility (Kaufman, 1975). If the third factor is viewed as a non-intellectual factor, then the construct validity of the WISC-R as an *intellectual* measure for different groups is supported even more strongly by these data.

Interpretation of the factor scores of VC and PO appears to be equally appropriate for all of the groups included in this study. Following Kaufman (1975), the Verbal Scale IQ can be used for the VC factor, and if M is substituted for Co, the Performance IQ can be used as for the PO factor. However, the FD factor for Chicanos should probably involve four subtests (adding PA to the usual three), and the FD factor scores should probably *not* be used with Blacks and Native-American Papagos unless other data suggest its existence and/or predictive utility for these groups.

Confidence in the appropriateness of the WISC-R as a measure of intellectual ability for different groups is increased by the fact that (a) a large general factor was clearly apparent in about the same form and amount for all groups. Thus, the usual interpretation of the Full Scale IQ as an index of general intelligence appears to be equally appropriate for Anglo and non-Anglo groups. Further, the Verbal-Performance scale distinction appears also to be equally appropriate for the four groups.

Finally, the conclusions of this study provide increased confidence in the construct validity of the WISC-R for different groups. Construct validity evidence is certainly a *necessary* but not a sufficient condition for fairness in test use. These conclusions, of course, do not reveal whether or not the common predictions and classifications based on the WISC-R are equally valid for diverse sociocultural groups.

References

- Cohen, J. A factor-analytically based rationale for the Wechsler Adult Intelligence Scale. *Journal of Consulting Psychology*, 1957, 21, 451-457.
- Cohen, J. The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *Journal of Consulting Psychology*, 1959, 23, 285-299.
- Kaufman, A. Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology*, 1975, 43, 135-147.
- Lindsey, J. *The factorial organization of intelligence in children as related to the variables of age, sex, and subculture*. (Doctoral dissertation, University of Georgia, 1967). *Dissertation Abstracts*, 1967, 27, 3664A-3665A. (University Microfilms No. 67-3567)
- Reschly, D., & Jipson, F. Ethnicity, geographic locale, age, sex, and urban-rural residence as variables in the prevalence of mild retardation. *American Journal of Mental Deficiency*, 1976, 81, 154-161.
- Sattler, J. *Assessment of children's intelligence*. Philadelphia, Pa.: Saunders, 1974.
- Semler, I., & Iscoe, I. Structure of intelligence in Negro and white children. *Journal of Educational Psychology*, 1966, 57, 326-336.
- Silverstein, A. An alternative factor analytic solution for Wechsler's intelligence scales. *Educational and Psychological Measurement*, 1969, 29, 763-767.
- Silverstein, A. Factor structure of the Wechsler Intelligence Scale for Children for three ethnic groups. *Journal of Educational Psychology*, 1973, 65, 408-410.
- Silverstein, A. Alternative factor analytic solutions for the Wechsler Intelligence Scale for Children-Revised. *Educational and Psychological Measurement*, 1977, 37, 121-124.
- Van Hagen, J., & Kaufman, A. Factor analysis of the WISC-R for a group of mentally retarded children and adolescents. *Journal of Consulting and Clinical Psychology*, 1975, 5, 661-667.

Received December 6, 1976 ■

General Versus Specific Traits in the Assessment of Anxiety

Martin Mellstrom, Jr., Marvin Zuckerman, and George A. Cicala
University of Delaware

The study compared the predictive validity of anxiety measures based on the general trait approach and approaches that include the situation dimension in the test. The subjects were 56 male and 58 female undergraduates. The subjects were pretested with general and specific trait anxiety measures and were later exposed to three situations involving a rat, a test, and social anxiety. The results indicated that the predictive validity of the specific measures was significantly greater than that of the general measures in 7 of 32 comparisons, whereas the reverse never occurred. Other analyses showed some support for the presence of behavioral consistency; for example, the generalizability coefficient for persons across situations and response modes was about .35, but little support was found for the traditional general trait tests that purport to measure them.

During the past decade there has been considerable debate in the field of personality and assessment on the relative importance of traits and situations. The traditional "trait" approach, which calls attention to enduring intrapsychic dispositions and deemphasizes the role of the environment, has been attacked on both theoretical and empirical grounds, and alternative models have been proposed. The behavioral or "situationist" approach (e.g., Kanfer & Saslow, 1965; Mischel, 1968), which has been one of the most widely accepted of these alternatives, can be characterized by an emphasis on the role of the current environment or situation in determining behavior. More recently, proponents of the interactional approach have suggested that neither trait nor situationist accounts of behavior are adequate, and that

the person and the situation must be considered in their interaction to provide an adequate conceptualization of behavior (cf. Bowers, 1973; Ekehammar, 1974; Endler & Magnusson, 1976).

Although reviews of the evidence on which of these approaches holds the most promise as a model for the study of personality (e.g., Bowers, 1973; Ekehammar, 1974; Endler & Magnusson, 1976) seem to support some form of interactional approach, it is not yet clear that an application of this approach to assessment problems will yield gains in our ability to understand and predict behavior. Studies of the problem in the area of anxiety (e.g., D'Zurilla, 1965; Hodges & Spielberger, 1966; Lamb, 1973; Mellstrom, Cicala, & Zuckerman, 1976; Paul, 1966) have compared the predictive validity of "anxiety trait" (A-Trait) tests that included the situation variable in the design of the test with those that did not. In these studies, "general A-Trait" tests, like that of Spielberger, Gorsuch, and Lushene (1970), which are intended to measure relatively stable individual differences in general anxiety proneness, are evaluated against "specific A-Trait" measures, which assess a person's disposition to be anxious in a particular situation or class of situations, for example, classroom examination situations. The idea of specific traits can be viewed

This article is based on a dissertation submitted by the first author in partial fulfillment of the requirements for the PhD degree from the University of Delaware. The authors would like to thank the following students for their work as experimenters in the present study: John Baltadonis, John Cavanaugh, Pamela Gravatt, Sandra Herman, George Laskaris, Frank Shields III, and Nancy Trump.

Martin Mellstrom, Jr., is now at the Psychology Service, Veterans Hospital, Lyons, New Jersey 07939. Requests for reprints should be sent to Marvin Zuckerman, Department of Psychology, University of Delaware, Newark, Delaware 19711.

as a redefinition of the trait concept so that it takes into account the situation and the Person \times Situation interaction.

Although the results, on the whole, have indicated that specific A-Trait measures are more predictive of anxiety in the corresponding criterion situation, there have been instances in which the general A-Trait tests were nearly or equally as predictive. This has seemed to happen when there was some threat to self-esteem ("ego threat") in the criterion situation (Mellstrom et al., 1976). Thus, the issue of whether anxiety assessment procedures based on a general trait model of personality are inferior to those that introduce the situation variable into assessment has been unresolved because the variable of ego threat has not been adequately explored and only a small number of anxiety situations have been evaluated. The present study attempted to help clarify this issue by having individuals who were pretested with general and specific A-Trait tests experience three anxiety situations, two of which posed a clear ego threat (social and test anxiety situations) and one of which did not (a rat situation).

Method

Subjects

Fifty-six male and 58 female students enrolled in introductory psychology at the University of Delaware were selected for the study. They participated to fulfill a course requirement.

Predictor Tests

Since all the predictors have been described in the Mellstrom et al. (1976) study, they will be only briefly described here.

Two of the pretests can be regarded as traditional A-Trait measures: (a) The trait form of the State-Trait Anxiety Inventory (STAI A-Trait; Spielberger et al., 1970) measures individual differences in anxiety proneness and has been found to correlate .73 with the Neuroticism scale of the Eysenck and Eysenck (1964) Personality Inventory (Mellstrom et al., 1976); (b) the Neuroticism (N) scale of the Eysenck Personality Inventory (EPI) was the second general anxiety measure used.

Two pretests provided indices of both general and specific A-Trait: (a) The Geer (1965) Fear Survey Schedule (FSS) consists of self-rated fear reactions to 51 commonly feared objects and situations. Subjects' total fear scores across all 51 items were used

as a general fearfulness or A-Trait measure, and their scores on the particular items were used as specific A-Trait measures. (b) Another test that provides indices of both general and specific A-Trait is a modified form of the Zuckerman Inventory of Personal Reactions (ZIPERS; Zuckerman, 1976, 1977). It consists of 12 situations in which subjects indicate the degree to which each situation elicits in them each of 13 reactions. Factor analysis of the response dimensions of the scale has yielded several factors including the fear arousal factor used in the present research. By summing a subject's scores across all situations on the responses comprising the fear factor, a measure of general A-Trait can be derived. A subject's score in response to any one of the 12 situations yields a specific A-Trait measure. The test was modified by adding social, test, and rat situations, which matched the real situations in the study, to the usual form of the test. Responses to these situations were not included in subjects' total scores, which were obtained by summing responses across the other situations.

Anxiety Situations

Rat. In the rat situation, an albino laboratory rat with clipped front incisors was housed in a wire cage located on a table in the corner of the 4 m square \times 3.5 m high experimental room.

Test. The test situation was similar to that used by Sarason (1957) to study test anxiety and involved a memory drum used to provide a serial verbal learning task. Ten nonsense syllables were presented at the rate of one every 1 sec, and the subject's task was to say when a given syllable appeared what the next one to appear would be. According to Sarason, when subjects are told that this task is a measure of intelligence, the resulting evaluation anxiety disrupts the performance of those who are "test anxious." Subjects spoke into a tape recorder connected to an audio oscillator so that a "beep" (inaudible to the subject) was recorded on the tape each time a new word was presented on the memory drum. Using this arrangement, subjects could be left alone to perform the task, and their tape can be scored later for the number of trials required to learn the list. Thus, anxiety in this situation would not be contaminated by social anxiety related to the presence of other people.

Social situation. This situation consisted of a non-directive interview during which subjects were asked to "talk about themselves" for 2 minutes. The first author conducted the interview while he and two other experimenters, one male and one female, gazed directly and continuously at the subject. This procedure was followed because of Marks's (1969) observation that one of the cardinal features of social anxiousness is a fear of the gaze of others. Although it never occurred, experimenters were instructed to stop gazing if it appeared that the subject was becoming very distressed by it. A videotape camera, focused on the subject, was also located in the room, but subjects were given no indication of whether it

was operating until the end of the experimental session. The situation could be observed from an adjoining room, where the videotape deck was located, through a one-way mirror.

Subjective Situational Anxiety Measures

The three tests used to measure subjective anxiety in each situation were the Anxiety State (A-State) scale of Spielberger et al.'s (1970) STAI; the A-State form of the ZIPERS, which consists of the items of the 13 reactions on the ZIPERS, with instructions to indicate the reactions being experienced "now"; and the fear thermometer (FT) of Walk (1956), in which subjects place a check mark on a 1-10 scale to indicate the amount of fear or anxiety being experienced at that moment.

Subjects were administered the predictor tests during the first 4 weeks of the semester. Six weeks later, each subject participated in each anxiety situation once, with the temporal order of the situations determined randomly for each subject. To minimize carryover effects from one situation to the next, subjects were scheduled for no more than one situation on a given day.

Rat Situation Procedure

Taped instructions directed the subject to approach the rat and lift it up. In this, as in the other two situations, subjects were told that they did not have to participate if they did not wish to, but no subject chose this option. Following the instructions, subjects were administered the first two A-State scales (STAI A-State and ZIPERS-State). While they completed these scales, the experimenter rated them on a modified form of Paul's (1966) Behavior Checklist in which the experimenter rates the degree to which each of the checklist behaviors is noticeable. An extra item of "overall anxiety" was added to the checklist.

The behavioral test was identical to the one used in the Mellstrom et al. (1976) snake situation and is basically a behavior approach test, yielding a "task score" and a latency measure as the behavioral fear indices. After this task was completed, the subject filled out a FT as he or she stood in the location of the closest approach to the rat. In all three situations, subjects were told that they did well relative to others to prevent possible loss of self-esteem.

Test Situation Procedure

Taped instructions told subjects that they would be asked to perform a difficult memory task that is an indicator of one's intelligence. Then the first two A-State scales were completed while the experimenter filled out the Behavior Checklist. Next, the memory task was begun, and the experimenter left the room. At the end of 15 minutes, he or she returned, terminated the task, and administered the FT.

Social Situation Procedure

Taped instructions asked subjects to talk about themselves for 2 minutes. The first two A-State scales were then administered, the "interview" was conducted, and the FT was completed. Before subjects left, they were told that they had been videotaped but that if they objected, the tapes would be erased. No subjects requested this. After subjects left, the experimenters watched the tape of the interview and completed the Behavior Checklist. The videotapes provided additional behavior ratings made by a separate set of observers, but these ratings were uncorrelated with the other measures in the study, so they will not be reported.

Results

Before the main results are discussed, it should be noted that there was a truncation of range in the behavioral measure of the rat situation (task score), since only 6 males and 15 females could not perform the task. As a consequence, the correlations of this variable with the predictors were expected to be somewhat low.

Comparative Validity of the General and Specific Predictors

In the previous study (Mellstrom et al., 1976), the correlations among the predictor measures indicated three fairly distinct

Table 1
Intercorrelations Among the Three Classes of Predictor Measures

Scale	Omnibus	Neuroticism	A-Trait	
			Rat	Test
Neuroticism	.49			
Rat A-Trait	.56	.28		
Test A-Trait	.80	.40	.46	
Social A-Trait	.73	.48	.43	.65

Note. Omnibus = summed standard scores of the total scores on the Geer Fear Survey Schedule and Zuckerman Inventory of Personal Reactions. Neuroticism = summed standard scores on the State-Trait Anxiety Inventory Trait scale (A-Trait) and Eysenck Personality Inventory Neuroticism scale. Rat A-Trait = summed standard scores of the rat A-Trait measures of the Geer Fear Survey Schedule and the Zuckerman Inventory of Personal Reactions. Test and social A-Trait composites were similarly derived. All correlations were significant at $p < .01$.

Table 2

Validity Coefficients for Three Classes of Predictors for Males

Situational anxiety measure	Predictor			<i>t</i> -test comparison		
	I: Omnibus	II: Neuroticism	III: Specific	I vs. II	I vs. III	II vs. III
Rat						
Self-report	.38**	.34**	.65**	<i>ns</i>	III > I	III > II
Observational rating	.36**	.13	.39**	<i>ns</i>	<i>ns</i>	<i>ns</i>
Behavioral	-.09	.06	.15	<i>ns</i>	<i>ns</i>	<i>ns</i>
Test						
Self-report	.52**	.38**	.48**	<i>ns</i>	<i>ns</i>	<i>ns</i>
Observational rating	.04	-.05	.02	<i>ns</i>	<i>ns</i>	<i>ns</i>
Behavioral measure	-.02	.20	.03	<i>ns</i>	<i>ns</i>	<i>ns</i>
Social						
Self-report	.33*	.40**	.44**	<i>ns</i>	<i>ns</i>	<i>ns</i>
Observational rating	.26*	.29*	.19	<i>ns</i>	<i>ns</i>	<i>ns</i>
% significant <i>r</i>	38	38	50			

Note. Omnibus = summed standard scores of the total scores on the Geer Fear Survey Schedule and Zuckerman Inventory of Personal Reactions. Neuroticism = summed standard scores on the State-Trait Anxiety Inventory Trait scale (A-Trait) and Eysenck Personality Inventory Neuroticism scale. Rat A-Trait = summed standard scores of the rat A-Trait measures of the Geer Fear Survey Schedule and the Zuckerman Inventory of Personal Reactions. Test and social A-Trait composites were similarly derived.

* $p < .05$.

** $p < .01$.

classes: neuroticism, omnibus, and specific A-Trait measures as represented by measures like the EPI-N, total score on the FSS, and a single item on the FSS, respectively. Measures within a given class were found to correlate more highly with each other than with other measures, and to show very similar patterns of correlation with the situational criteria. Therefore, to simplify interpretation of the data, tests comprising one class were combined to form one composite measure by summing subjects' standard scores on each test. For example, subjects' standard scores on the STAI A-Trait and EPI-N scales were summed to provide a composite labeled *neuroticism*. An additional justification for this procedure is that the focus of the study was not on the validity of any single test, which may be affected by the amount of effort put into its construction, but on the validity of different strategies of assessment as represented by more than one measure.

The response data of the present study were prepared for analysis in the same way. On the basis of an a priori decision, based on the foregoing theoretical rationale, sub-

jects' standardized total scores on the FSS and ZIPERS were summed to provide a composite labeled *omnibus*. The STAI A-Trait and EPI-N scales were combined to form a neuroticism composite, and the specific A-Trait measures derived from the ZIPERS and FSS were similarly combined for each specific disposition, yielding specific composites for "social anxiety," "test anxiety," and "rat anxiety."

As in the prior study, situational criteria of the same type were combined to form self-report, observers' rating, and behavioral index composites. STAI A-State, ZIPERS-State, and the FT comprised the "self-report" composite, total score on the Behavior Checklist and the additional "overall anxiety" item comprised the "observers' rating" composite, and, in the rat situation, the task score and latency formed a "behavioral index" composite.¹

¹ Anyone interested in the correlations between the single predictor measures and the separate fear response measures can obtain a table of these correlations from the second author.

Table 1 shows the correlations among the composite measures representing the several classes of predictors for the entire sample of subjects. The omnibus composite showed moderate to strong correlations, ranging from .49 to .80 with all the other predictors, whereas the neuroticism composite showed somewhat weaker but significant correlations. The specific A-Trait composites for test and social anxiety correlated highly (.65) and were more highly related to the neuroticism composite than was the rat A-Trait composite.

Tables 2 and 3 show the correlations between the three classes of predictors and the situational fear measures for the male and female samples, respectively. The validity coefficients of the specific compositions represent the relations between the situational anxiety measures and the specific composite designed to predict that anxiety. To compare the validity coefficients of the three types of A-Trait predictors, a *t* test for the significance of the difference between two correlation coefficients in correlated samples (Ferguson, 1971) was used. The right-hand por-

tion of the tables shows the results of these comparisons.

For the rat situation, there were several instances in which the specific composite had significantly more accuracy in the prediction of criteria than either the omnibus or neuroticism composites. Also, the specific composite correlated significantly with all three types of situational measures except the behavioral criterion for the males, which was limited by range restriction, whereas the neuroticism composite correlated significantly with self-reports and observer's ratings but not with behavioral measures.

In the test situation, there was little or no evidence of differential predictive power for the three types of predictors. For the females in the social situation (Table 3), the specific measure was significantly better than the omnibus one in predicting both types of situational responses. Overall, the specific measures were significantly more predictive in 7 of 32 comparisons, whereas the reverse never occurred; that is, the general measures were never significantly more predictive. Also, the

Table 3
Validity Coefficients of Three Classes of Predictors for Females

Situational anxiety measure	Predictor			<i>t</i> -test comparison		
	I: Omnibus	II: Neuroticism	III: Specific	I vs. II	I vs. III	II vs. III
Rat						
Self-report	.25	.34**	.52**	<i>ns</i>	III > I	<i>ns</i>
Observational rating	.27*	.18	.41**	<i>ns</i>	<i>ns</i>	<i>ns</i>
Behavioral	.02	.02	.34**	<i>ns</i>	III > I	III > II
Test						
Self-report	.54**	.51**	.50**	<i>ns</i>	<i>ns</i>	<i>ns</i>
Observational rating	.16	-.07	.23	<i>ns</i>	<i>ns</i>	<i>ns</i>
Behavioral	.07	.15	.08	<i>ns</i>	<i>ns</i>	<i>ns</i>
Social						
Self-report	.49**	.51**	.66**	<i>ns</i>	III > I	<i>ns</i>
Observational rating	.19	.26*	.38**	<i>ns</i>	III > I	<i>ns</i>
% significant <i>r</i>	25	38	75			

Note. Omnibus = summed standard scores of the total scores on the Geer Fear Survey Schedule and Zuckerman Inventory of Personal Reactions. Neuroticism = summed standard scores on the State-Trait Anxiety Inventory Trait scale (A-Trait) and Eysenck Personality Inventory Neuroticism scale. Rat A-Trait = summed standard scores of the rat A-Trait measures of the Geer Fear Survey Schedule and the Zuckerman Inventory of Personal Reactions. Test and social A-Trait composites were similarly derived.

* $p < .05$.

** $p < .01$.

percentages of significant ($p < .01$) correlations for the omnibus, neuroticism, and specific predictors (Tables 2 and 3) were 38%, 38%, and 50%, respectively, for the males and 25%, 38%, and 75% for the females. The results suggest that the specific predictors had somewhat greater overall predictive accuracy than the other two types of A-Trait measures.

Proportions of Variance: Persons Versus Situations

To assess the relative importance of persons, situations, and their interaction, the ZIPERS data were subjected to three-way analyses of variance. In the first analysis, each subject represented a level of the person source, whereas the social, test, and rat situations described in the inventory represented levels of the situation source. The three response items that represented the mode of response source were "Your heart beats faster," "You feel fearful," and "You get out of the situation or avoid it." The rationale for using these three responses was that each represents one of the three basic human response channels: physiological, phenomenal, and behavioral. Also, it was thought preferable to have the same number of response modes as situations.

After the analyses of variance were performed, the proportions of variance explained

Table 4
Percentages of Variance Accounted for by Sources, with Individuals Comprising Levels of the Person Source

Source	Type of situation	
	Hypo- thetical	Actual
Person (P)	35.2	29.0
Situation (S)	.0	.6
Response mode (R)	4.6	5.7
P \times S	14.0	21.5
P \times R	16.8	11.7
S \times R	2.5	.7
Residual	26.8	30.8

Note. The hypothetical situations were the rat, test, and social situations described in the Zuckerman Inventory of Personal Reactions. The actual situations were those actually experienced by subjects in the study.

Table 5
Correlations of Responses Across Actual Situations for Each Measure

Situation	Test	Social
STAI State scale		
Rat	.34**	.45**
Test		.39**
ZIPERS State scale		
Rat	.45**	.37**
Test		.51**
Fear Thermometer		
Rat	.22*	.27**
Test		.40**
Behavior Checklist composite		
Rat	.24*	.27**
Test		.16

Note. STAI = State-Trait Anxiety Inventory; ZIPERS = Zuckerman Inventory of Personal Reactions.

* $p < .05$.

** $p < .01$.

by each source were calculated, following guidelines described by Endler (1966). The proportions were calculated twice under a mixed model (persons random, situations, and response modes fixed), once assuming a minimum and once a maximum triple interaction. Since the two sets of results were quite similar, it was possible to make the assumption of a zero triple interaction, permitting a unique solution for each component.

Regarding statistical significance, this $n = 1$ model permits F tests on only the situation and response mode main effects and the Situation \times Response Mode interaction effect. Of these three, only the situation effect was not significant in either the hypothetical or actual situation data sets.

Table 4 shows that the person source accounted for 35% and 29% of the variance in the responses to the hypothetical and actual situations, respectively. In contrast, situations accounted for only 0%–1%, a finding related to the fact that the means of the three situations were nearly identical. Mode of response explained only 5%–6% of the responses to the hypothetical and actual situations, and

the Person \times Situation interaction accounted for 14% or 22%, depending on the data set. The residual accounted for 27% or 31%, depending on the data set. Females had significantly higher anxiety levels than the males on the ZIPERS specific A-Trait predictors, but they did not differ from males on any of the A-State responses in the actual situations.

To assess the importance of individual differences when defined by traditional general A-Trait tests, more analyses of variance were performed on the ZIPERS data when the person source was represented by groups of subjects scoring high, medium, and low on the neuroticism composite. The situation and response sources were the same as before, and the analysis was done on both the hypothetical and actual situation data sets.

The results of this change were that the A-Trait or neuroticism source explained only 12% and 9% for the hypothetical and actual situation data sets, respectively. As before, situation explained little or no variance, but in this analysis the error term accounted for 69%–81% of the variance.

Individual Consistency Across Situations

As pointed out by Endler and Magnusson (1976), one of the main postulates of the trait position is that the rank ordering of individuals on a given trait should remain the same from situation to situation. To determine the validity of this postulate for the data of the present study, the correlations of a given situational anxiety measure with the same measure taken in other situations were examined. Table 5 shows that the size of the correlations varied considerably from measure to measure, ranging from .16 to .51, with the median correlation at about .36.

Another index of consistency, the generalizability coefficient (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), was computed on the ZIPERS data and submitted to analysis of variance. The person source had the largest coefficients, .38 and .31, respectively, for the hypothetical and actual situations, whereas the situation source had coefficients near zero, demonstrating no consistent or generalizable effect across persons.

Discussion

The findings on the comparative validity of the general and specific A-Trait measures corroborate those of the earlier study (Mellstrom et al., 1976) and suggest that the specific measures may be more useful for prediction of fear responses. The specific composites surpassed the general ones in the rat situation for both sexes and in the social situation for females. In both of these instances, it would seem that subjects were capable of imagining how they would respond in the situation, and that this estimate was more accurate than one based on a measure of the person's general anxiousness.

Instances in which the general tests equaled the predictive accuracy of the specific ones (i.e., in the social situation for the males and in the test situation for both sexes) can be explained in several ways. First, the predictive power of the specific predictors may have been limited by two factors: (a) Subjects probably had no prior experience with anything similar to the test situation of the study, making it difficult for them to predict on the ZIPERS how they would react. and (b) There was a lack of correspondence between the test "situation" described in the FSS test item and the actual situation confronted; that is, performing a serial verbal learning task in an experiment differs considerably from "failing a test," the situation described in the FSS test item. Second, the specific predictors did not surpass the general ones in predicting the self-report criteria of the test situation, because the general ones achieved relatively high levels of predictive accuracy for this situation. It may be that the ego threat present in such situations is the main cause of anxiety, permitting the neuroticism composite to be predictive of these self-reports. Third, the general and specific predictors may have shown equal predictive validity for the males in the social situation because many males may be less likely than females to accurately imagine their responding in social situations. That is, in line with traditional sex roles, females may be more attentive to subtle social cues and to their own responding in such situations.

Thus, while the idea of including the situa-

tion variable in assessment seems to be theoretically sound and often yields increments in predictive validity, variables like the "ego involvement" of subjects, their prior experience (or lack of it) with situations like the ones of interest, and the similarity between the criterion situation and the situation described in the specific measure may attenuate the validity of the specific measures.

It should be noted that the rat situation may have aroused ego threat in addition to rat fear, since many males, and perhaps females as well, may consider it childish or foolish to show fear in such a situation, whether the experimenter is female or male. This may explain the significant correlations between the neuroticism composite and the self-report criterion in this situation. Although the effect of mixed-sex experimenter-subject dyads in this situation, as well as in the other situations, is not known, it seems reasonable to assume, without evidence to the contrary, that subjects would feel equally threatened by potential failure in the presence of a same-sex peer as they would in front of one of the opposite sex.

The findings on the consistency and person-versus-situation issues provided some support for the idea of behavioral continuity (cf. Block, 1971). The indices of consistency suggest that persons' anxiety levels were somewhat generalizable across situations. In addition, analyses of variance on the Mellstrom et al. (1976) data, presented in Zuckerman and Mellstrom (1977), also showed this pattern; the person source explained about 29% of the variance. However, in both studies, when the person source was defined by levels of the traditional A-Trait measures, the percentages of explained variance dropped. In the present study, they dropped from 35% to 12% and 29% to 9% for the hypothetical and actual situation data sets, respectively.

Thus, although some degree of individual consistency of response may exist, we found little predictive power for measures of the trait that is supposed to underlie the consistency. The most likely explanation is that traditional A-Trait tests are not valid measures of the construct that they purport to measure. Alternately, the trait may be more

consistent in some people than in others; that is, some people may be consistent while others are "interactive," being differentially responsive to situations (cf. Bem & Allen, 1974). The specific measures should be predictive for both types of persons as long as the situation is similar to one the person has reacted to on past occasions.

One problem that remains to be resolved is the optimal degree of specificity for prediction. Too broad a definition of situations in the tests may attenuate the predictive potential of the tests, whereas too narrow a definition may limit their usefulness to extremely narrowly defined situations.

References

- Bem, D. J., & Allen, A. On predicting some of the people some of the time. *Psychological Review*, 1974, 81, 506-520.
- Block, J. *Lives through time*. Berkeley, Calif.: Bancroft Books, 1971.
- Bowers, K. S. Situationism in psychology: An analysis and a critique. *Psychological Review*, 1973, 80, 307-336.
- Cronbach, L., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.
- D'Zurilla, T. J. Effects of behavioral influence techniques applied in group discussion on subsequent verbal participation. (Doctoral dissertation, University of Illinois, 1964). *Dissertation Abstracts*, 1965, 25, 4813A-4814A. (University Microfilms No. 65-804)
- Ekchammar, B. Interactionism in personality from a historical perspective. *Psychological Bulletin*, 1974, 12, 1026-1048.
- Endler, N. S. Estimating variance components from mean squares for random and mixed effects analysis of variance models. *Perceptual and Motor Skills*, 1966, 22, 559-570.
- Endler, N. S., & Magnusson, D. Toward an interactional psychology of personality. *Psychological Bulletin*, 1976, 83, 956-974.
- Eysenck, H. J., & Eysenck, S. B. An improved short questionnaire for the measurement of extraversion and neuroticism. *Life Sciences*, 1964, 3(10), 1103-1109.
- Ferguson, G. A. *Statistical analysis in psychology and education*. New York: McGraw-Hill, 1971.
- Geer, J. H. The development of a scale to measure fear. *Behaviour Research and Therapy*, 1965, 3, 45-53.
- Hodges, W. F., & Spielberger, C. D. The effects of threat of shock on heart rate for subjects who differ in manifest anxiety and fear of shock. *Psychophysiology*, 1966, 2, 287-294.

- Santer, F. H., & Saslow, G. Behavioral analysis: An alternative to diagnostic classification. *Archives of General Psychiatry*, 1965, 12, 529-538.
- Lamb, D. H. The effects of two stressors on state anxiety for students who differ in trait anxiety. *Journal of Research in Personality*, 1973, 7, 116-126.
- Marks, I. M. *Fears and phobias*. New York: Academic Press, 1969.
- Mellstrom, M., Cicala, G. A., & Zuckerman, M. General versus specific trait anxiety measures in the prediction of fear of snakes, heights, and darkness. *Journal of Consulting and Clinical Psychology*, 1976, 44, 83-91.
- Mischel, W. *Personality and assessment*. New York: Wiley, 1968.
- Paul, G. *Insight vs. desensitization in psychotherapy*. Stanford, Calif.: Stanford University Press, 1966.
- Sarason, I. G. Test anxiety, general anxiety, and intellectual performance. *Journal of Consulting Psychology*, 1957, 21, 485-490.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *STAI manual for the State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Walk, R. D. Self-ratings of fear in a fear-invoking situation. *Journal of Abnormal and Social Psychology*, 1956, 52, 171-178.
- Zuckerman, M. General and situation-specific traits and states: New approaches to assessment of anxiety and other constructs. In M. Zuckerman & C. D. Spielberger (Eds.), *Emotions and anxiety: New concepts, methods and applications*. Hillsdale, N.J.: Erlbaum, 1976.
- Zuckerman, M. Development of a situation-specific trait-state test for the prediction and measurement of affective responses. *Journal of Consulting and Clinical Psychology*, 1977, 45, 513-523.
- Zuckerman, M., & Mellstrom, M., Jr. General vs. specific traits as predictors of self-reported states and behavior in real situations. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, N.J.: Erlbaum, 1977.

Received January 20, 1977 ■

INSTRUCTIONS TO AUTHORS

Style of manuscripts. Manuscripts must be prepared in the style described in the *Publishing Manual of the American Psychological Association* (2nd ed.). Instruction on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors should submit manuscripts in triplicate and should keep a copy to guard against loss.

Abstracts. Manuscripts of regular articles must be accompanied by an abstract of 100-175 words. Manuscripts of brief reports must be accompanied by an abstract of 75-100 words. All abstracts must be typed on a separate sheet of paper.

Brief Reports. The *Journal of Consulting and Clinical Psychology* will accept Brief Reports of search studies in clinical psychology. The procedure is intended to permit the publication of soundly designed studies of specialized interest or limited importance which cannot now be accepted as regular articles because of lack of space. Several pages in each issue will be devoted to Brief Reports, published in the order of their receipt without respect to the dates of receipt of the regular articles. Most Brief Reports appear in the first or second issue to go to press following their final acceptance.

An author who wishes to submit a Brief Report:

1. Sends the Brief Report, limited to two printed pages and prepared according to the specifications given below.

2. Agrees not to submit the full report to another journal of general circulation.

Specifications. The Brief Report should give a clear, condensed summary of the procedure of the study and as full an account of the results as space permits.

To ensure that a brief report does not exceed two printed pages, follow these instructions for typing: (a) Set typewriter to a 48-space line. (b) Type text. (c) Count all lines except abstracts, title, and by-line, including acknowledgments. If you have exceeded 185 lines, shorten the material. (d) If your brief report barely exceeds 75 lines (one printed page), try to edit to 75 lines exactly. Reports slightly exceeding one page will be edited for length in order to use space economically.

In Brief Reports, headings, tables, and references are avoided or, if essential, must be counted in the 75 lines. This journal no longer requires an extended report, but, if one is available, each Brief Report must be accompanied by the following footnote:

Requests for reprints and for an extended report of this study should be sent to John Doe (give the author's full name and address).

The footnote should be typed on a separate sheet and not counted in the 75-line quota.

Effect of Subliminal Stimulation of Symbiotic Fantasies on Behavior Modification Treatment of Obesity

Lloyd H. Silverman, April Martin, Roseann Ungaro, and
Eric Mendelsohn

New York Veterans Administration Regional Office, New York, and
Research Center for Mental Health, New York University

In two studies, obese women were treated in a behavior modification program for overeating, in Study 1 for 8 weeks and in Study 2 for 12 weeks. In both studies, the behavior programs were accompanied by subliminal stimulation, with half of the subjects receiving the verbal message MOMMY AND I ARE ONE, intended to stimulate symbiotic gratification fantasies, and the other half, a control message. Weight loss was measured at the end of the program and at follow-up times; in Study 1, 4 weeks after termination and in Study 2, at 4 and 12 weeks posttermination. In both studies the symbiotic condition gave evidence of enhancing weight loss, though it was only at follow-up that the difference between the groups attained statistical significance. This finding, when viewed in conjunction with results from earlier studies of schizophrenics and insect phobics, supports the proposition that the subliminal stimulation of symbiotic fantasies can enhance the effectiveness of therapeutic interventions of various kinds.

During the past 12 years, a research method termed *subliminal psychodynamic activation* has been used in the experimental study of a critical aspect of psychoanalytic theory: the relationship between psychopathological behavior and unconscious libidinal and aggressive wishes. In most of this work, the stimuli were designed to stir up these wishes with the prediction that their subliminal presentation (as compared to the subliminal presentation of relatively neutral stimuli) would measurably intensify particular kinds of psychopathology. In over 20 studies completed to date (summarized in Silverman, 1976),

this expectation has been borne out. The subliminal presentation of a wish-related stimulus produced pathological reactions that did not appear after the subliminal presentation of a neutral stimulus; and in a number of studies, they also did not appear after the wish-related stimulus was presented supraliminally and in the subject's awareness.¹

In another aspect of the research, however, instead of using stimuli designed to *stir up* unconscious wishes and intensify psychopathology, a stimulus was used that was intended

We wish to thank Deborah Brenner, Jill Kline, and Susan Packer for their participation as behavior modification interventionists.

April Martin is now at the Institute for Contemporary Psychotherapy, New York City.

Eric Mendelsohn is now at New York Hospital, Westchester Division.

Requests for reprints, copies of the stimuli used, and other information relevant to replication should be sent to Lloyd Silverman, Research Center for Mental Health, New York University, 6 Washington Place, Room 450, New York, New York 10003.

¹ These studies should be distinguished from the more traditional experiments in the "subliminal area" that have aimed at "finding" the tachistoscopically exposed stimulus (in transformed guise) in the subsequent productions of the subject, rather than at observing its pathological effects. For the reader whose contact with the subliminal area ended with the early skeptical critiques of the phenomenon, see the recent exhaustive and detailed review of Dixon (1971). Finally, for a discussion of why the supraliminal presentation of the same stimuli usually fails to trigger a pathological reaction and why subliminal presentation is then the method of choice for the laboratory study of psychodynamic aspects of psychoanalytic theory, see Silverman (1972).

safely gratify a particular wish and reduce ego pathology. This stimulus was the verbal message MOMMY AND I ARE ONE. Its use was based on the following two interrelated assumptions: (a) The fantasized gratification of the wish for oneness with mommy—the good mother of infancy—can ameliorate ego pathology of various kinds (cf. Silverman, Note 1); and (b) the subliminal presentation of the words MOMMY AND I ARE ONE has the power to activate this fantasy.

In support of the above assumptions, the following can be cited: First, in experiments carried out with eight groups of male schizophrenics (Bronstein, 1976; Kaplan, 1976; Kaye, 1975; Leiter, 1973; Silverman & Cancell, 1970; Silverman, Spiro, Weissberg, & Cancell, 1969; Spiro, 1975), the subliminal presentation of this "symbiotic gratification stimulus" when compared with the (subliminal) effects of a neutral control message has been found to reduce the degree to which ego pathology" is manifested within a laboratory session. Second, with two groups of male homosexuals (research volunteers), another ameliorative effect has been found—a decrease in anxiety and defensiveness within a laboratory session after the subliminal exposure of this same symbiotic gratification message (Silverman, Kwawer, Wolitzky, & Bron, 1973).

In addition to the above findings, there have been three studies in which more than a laboratory effect" has been demonstrated. In one, (Silverman, Frank, & Dachinger, 1974), the effectiveness of the symbiotic gratification stimulus as an aide in the behavioral treatment of insect phobias was demonstrated. Twenty women with insect phobias were seen twice weekly for six sessions. The first and last sessions were for pretreatment and post-treatment assessments of the degree of phobia, with the intervening four sessions for treatment—a variant of systematic desensitization. At each treatment session, subjects viewed scenes of insects that they had previously arranged hierarchically for their anxiety-arousing effects. Subjects began with the least fearful image and progressed to more frightening scenes. After each image, the subjects gave a subjective rating of the degree

of discomfort that they experienced. When discomfort ratings exceeded a specified level, the subjects looked into the tachistoscope for subliminal stimulation. Stimulation was repeated until the discomfort ratings for a particular image were below the criterion level, and subjects then progressed to the next image in the hierarchy. In the usual systematic desensitization paradigm, the imaging is accompanied by deep muscle relaxation. This study substituted subliminal stimulation for the muscle relaxation technique, with the experimental group receiving MOMMY AND I ARE ONE while subjects in the control group received the stimulus PEOPLE WALKING, intended as a (relatively) neutral verbal message. Following the four intervention sessions, on measures of both avoidance and anxiety, the experimental subjects showed a significantly greater degree of improvement of their phobic symptoms than the control subjects.

A second experiment (Silverman, Levinson, Mendelsohn, Ungaro, & Bronstein, 1975) investigated the effects of stimulating symbiotic fantasies during brief therapy with recently hospitalized male schizophrenics. Forty subjects were seen individually, three times a week, over a 6-week period. Treatment consisted of a "fantasy expression" procedure, in which the subjects were shown pictures and were encouraged to fantasize about them, with special emphasis on deriving pleasure from the fantasy and stressing the distinction between fantasy and reality. Half of the subjects were subliminally stimulated with MOMMY AND I ARE ONE several times during each fantasy expression session. The other half received as a control the stimulus PEOPLE ARE WALKING. Pretreatment and posttreatment assessments were made of "ego pathology" on the basis of cognitive and projective tests, interview ratings, and ratings of ward behavior made by the nursing staff. Both groups showed a reduction in the amount of ego pathology in evidence after treatment, but those who had received the experimental stimulus showed a significantly greater reduction.

In the third study (Parker, 1977), two groups of male and female college undergraduates ($n = 20$ in each group), matched

for academic performance, were given tachistoscopic stimulation at the beginning of a class four times a week over a 6-week summer term. For one group the stimulus was MOMMY AND I ARE ONE, whereas for their matched counterparts it was PEOPLE ARE WALKING. The students in the former group received grades on their final exam ("blindly" marked) that were significantly and substantially higher than did the controls (average marks of 90.4% and 82.7%, respectively).²

The present study was intended as an extension of this earlier work with a new subject population—obese women—and accompanying a different intervention—behavior modification training in weight control. It was hypothesized that for persons being treated for obesity with this form of therapy, those whose treatment was accompanied by the subliminal presentation of the symbiotic stimulus would lose more weight than similar persons in the same treatment who were presented with subliminal neutral stimulation. In addition to the point cited earlier about the general ameliorative effects of symbiotic gratification fantasies, the following specific rationale can be offered for predicting such a finding with this population: From psychoanalytic clinical observations of obese patients (Bruch, 1973; Bychowski, 1950), it can be inferred that overeating for the obese person is often motivated by ungratified unconscious wishes for a symbiotic experience. Thus, the fantasied gratification provided by repeated subliminal exposure to the MOMMY AND I ARE ONE stimulus was expected to make their overeating less necessary and thus aid them in successfully using the weight control program.³

Study 1⁴

Method

Subjects. Thirty obese women were recruited through an advertisement in a local newspaper. All subjects were at least 15% overweight, based on the 1959 Metropolitan Life Insurance norms for desirable weights for women (U.S. Department of Health, Education, and Welfare, 1967). Percentage of overweight was determined using the middle weight of the range given for a woman of medium frame at a given height as a baseline. To be eligible for the study, each subject had to state that she felt herself to be an overeater, that she was not currently involved in any organized program of treatment for

obesity, and that she had the time and motivation to attend the treatment sessions. In addition, a brief interview was conducted with each potential subject to screen out psychotic and borderline-psychotic applicants. Four women were excluded on this basis. During the course of data collection, three subjects dropped out after the first session (two of them had been assigned to the experimental group and one to the control), and they were replaced by other applicants.

Subjects were randomly assigned to an experimental and control group. The overall sample ranged in age from 22 to 59, with mean age of 30.7 for the

² In two other recent studies, subliminal symbiotic stimulation has been used with college students. Sackeim (1977) found that within a laboratory session, the MOMMY AND I ARE ONE stimulus heightened self-esteem (as measured by a semantic differential scale). On the other hand, Condon (1976) obtained negative results in attempting to replicate the findings of Silverman, Frank, and Dachinger (1974). Here, it may be important that the population Condon used, unlike the original population, did not consist of persons seeking treatment for their phobias. Instead, the sample was comprised of students who although manifesting a certain degree of phobic symptomatology, entered the study to fulfill a psychology class requirement. It thus may be that subliminal symbiotic stimulation to enhance the effectiveness of a treatment intervention, individuals must be motivated to overcome whatever behavior the treatment is intended to address. Further research is planned to test out this and other possibilities in accord with Condon's nonreplication.

³ The "wishes for a symbiotic experience" referred to above can be seen as related to the "symbiotic phase of development" (cf. Mahler, Pine, & Bergman, 1975), defined as the period of infancy when differentiation from mother and a sense of separateness from her are minimal and most incomplete. The "oneness" with her at this time can serve a number of needs: her presence is guaranteed; her "omnipotence" is shared; nurturance is always available; and she can offer both protection against external dangers and assistance in mastering internal dangers—that is, helping to control unacceptable impulses of various kinds. A legacy of this symbiotic phase of development is the merging wishes referred to above that are viewed as characterizing, to varying degrees, different people throughout their lives. That is, to the extent that needs for protection, omnipotence, nurturance, and so on, are powerful and still sought in the manner of early infancy, wishes for oneness with "mommy" arise. The extent to which they are then gratified depends on the degree of internal conflict such wishes generate as well as on external circumstances.

⁴ This study was part of a doctoral dissertation submitted by the second author (Martin, 1975) in partial fulfillment of the requirements for a doctoral degree at New York University.

experimental subjects and 31.8 for the control subjects. The percent overweight for the sample ranged from 15% to 94%, with means of 40.4% for the experimental group and 42.7% for the control. Neither of these differences approached significance, nor did the difference between the groups in racial composition (93% white). The subjects were asked to estimate the length of time that they have been overweight. The mean number of years estimated for the experimental group was 18.3 and for the control group 19.0, a difference that proved to be nonsignificant ($t < 1$).

Stimuli and tachistoscope. The symbiotic gratification stimulus consisted of the verbal message MOMMY AND I ARE ONE, printed in ink on a white 3×5 card in capital letters with the words MOMMY AND I on one line and the words ARE ONE on a second line. Another card contained the neutral control message PEOPLE WALKING presented on one line. The stimuli were shown through an electronically controlled mirror tachistoscope. The subject looked through an eyepiece at a blank field, and the stimulus was exposed from a second field. The viewing distance was 4 inches (1.3 m), and the surface brightness of a white card for the intensity settings of both fields was 5 ftL (17.1 cd/m²). Exposure time for the stimulus was 4 msec. In previous experiments under these conditions, no subject was able to recognize the content aspects of any stimulus, and less than 10% could discriminate between flashes of light produced by different stimuli (cf. Silverman, 1976).

Procedure. Subjects were randomly assigned to the experimental (symbiosis) or the control group. Two "interventionists," graduate students in psychology, conducted the behavioral treatment sessions, each working with an equal number of subjects in each stimulus group. The interventionists identified the stimuli by code letters appearing on the back of the stimulus cards and remained "blind" throughout the data collection as to which group each subject was in.

Each subject met with her interventionist individually, once a week for 8 consecutive weeks. At each session the subject's weight, in indoor clothing without shoes, was measured using a conventionalathroom scale. Treatment sessions were $\frac{1}{2}$ hour long. The 8-week program was designed to be similar to the behavior modification treatment program for obesity used by Wollersheim (1970). Subjects were instructed on how to keep records of the food they ate and its caloric content, how to systematically reduce the number of situations in which they ate, how to eat more slowly and with more awareness of sensation, and how to reward themselves for appropriate eating behavior. (The structure of the program and the specific techniques used are described more fully in Martin, 1975.)

At the beginning and end of each treatment session, the subject was instructed to look into the tachistoscope for a presentation of the subliminal stimulus. The first presentation was introduced in the following manner: The subject was asked to imagine herself in a situation in which she felt

tempted to overeat. She was asked to describe this situation fully—the place, time, circumstances, the food she was craving, and so on, until she reported that the image was very vivid. For example, she might describe her feeling when seeing cupcakes in the bakery window as she passed it. The experimenter would then say:

People usually feel a kind of tension when they want to eat something but are trying to tell themselves to resist it because they want to lose weight. This is a machine [the tachistoscope] which presents flashes of light. Researchers have found that these flashes can be useful in helping people to relax. If you are able to relax yourself at moments when you are craving something to eat which you know you shouldn't, you will be able to make a calmer decision to resist it. So I want you to look into the viewer, holding that image of the cupcakes trying to say "no" to them. I will say, "Ready, get set," and then you will see a flash of light. I will repeat this for a second flash a few seconds later. I want you to use these flashes to help you relax and walk calmly away from the bakery window.

The subject also was instructed that outside of the treatment sessions, whenever she found herself about to overeat, she should form a mental image of the flash of light she had seen in the machine and then try to refrain from eating. She was told that this would get easier when she had more experience with the flashes and that she would have a chance to see them again at the end of the session and at the beginning and end of every subsequent session. Eventually, it was explained, she would find that as she reached for food in an inappropriate way, she would automatically remember the flash of light and be able to remind herself to return to appropriate eating. The rationale for presenting the subliminal stimulus in this manner was that it served to arouse the subject's tension concerning eating behavior, which then could be reduced by the symbiotic fantasies that were being tachistoscopically activated. Thus, it was analogous to the procedure used in the desensitization study earlier described (Silverman et al., 1974).

Four weeks after the program ended (i.e., 12 weeks after the first session), the subjects returned for a follow-up weigh-in and were debriefed. The debriefing consisted of explaining the rationale for the study, showing both stimuli that had been used and informing each subject of the stimulus to which she had been exposed. Without exception, the subjects expressed surprise and in many instances disbelief that any stimulus had been exposed, insisting that they consciously saw nothing more than flickers of light during the experiment.

Results

Table 1 presents the mean weights of both groups initially, at the end of treatment (8th

Table 1
Means and Standard Deviations of Weight in Pounds for Study 1

Time	Experimental group		Control group	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Pretreatment	181.8	24.6	183.8	40.6
Posttreatment (8th week)	173.8	27.0	178.5	40.7
Follow-up (12th week)	170.9	27.2	179.4	40.9

Note. 1 pound = .4536 kg.

week), and at follow-up (12 weeks). Analyses were carried out in which the initial weights were covaried out of the 8th and 12th week weights.⁵ At the end of treatment, the results, although in the hypothesized direction, were not significant, $F(1, 27) = 1.84$, $p < .18$. At follow-up, however, the difference between the two groups was significant, $F(1, 27) = 7.08$, $p < .01$. As Table 1 indicates, significance was obtained at the latter time because during the 4-week follow-up period, the two groups behaved differently. The experimental group, on the average, continued to lose weight while the control subjects gained. Since the subjects had not been debriefed prior to the follow-up weigh-in, the difference in their eating behavior during the follow-up period can be ascribed to the continuing effects of the differential subliminal stimulation.

Table 2 presents the analogous findings for percent overweight. These results closely parallel the results for actual weight, with the difference between the two groups approaching significance after 8 weeks, $F(1, 27) =$

Table 2
Means and Standard Deviations of Percent Overweight for Study 1

Time	Experimental group		Control group	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Pretreatment	40.4	20.1	42.7	26.6
Posttreatment (8th week)	34.1	31.9	39.8	26.9
Follow-up (12th week)	32.0	19.9	40.5	27.7

3.75, $p < .06$, and reaching significance after 12 weeks, $F(1, 27) = 9.77$, $p < .004$.

Study 2

This study was intended as a replication and extension of Study 1, with several small changes and additions instituted.

Method

Subjects. Subjects again were recruited through newspaper advertisements. The criteria for selection were the same as in the initial study, except that a more extensive procedure was used for screening out psychotic and borderline-psychotic subjects. Rorschach and figure-drawing protocols were collected in an initial intake assessment, which together with the impression made by the subjects in a brief interview were reviewed by one of the authors (LHS) who has had extensive clinical experience. Six subjects were eliminated on this basis. Of the subjects who began the program, 11 dropped out (5 from the experimental and 6 from the control group) and were replaced. The final sample consisted of 13 subjects in each of the two groups, with mean ages of 31.7 and 36.1 (ages ranged between 22 and 57) and mean percent overweight of 37.9 and 37.4 (percent ranged between 15 and 118) for the experimental and control groups, respectively.⁶ Neither of these differences approached significance, nor did differences between the groups in racial composition (92% white) or in the number of years overweight (mean experimental = 20.1 and control = 21.0).

Stimuli and tachistoscope. The tachistoscopic conditions were identical to those of the initial experiment. However, the control stimulus was slightly altered so that it now consisted of the words PEOPLE

⁵ It is to be noted in Table 1 that with regard to the initial weights of the two groups, although their means were comparable, the standard deviation for the control group was considerably larger than the standard deviation for the experimental group. However, an analysis of covariance still could be carried out, since a test for homogeneity of regression for the two groups revealed that homogeneity was in evidence. Furthermore, with regard to the comparability of the experimental and control groups, for percent overweight (to be discussed) the standard deviations for the two groups were not discrepant.

⁶ In this study (in contrast to Study 1), percent overweight was calculated taking into account the subject's frame (designated as "small," "medium," or "large"). Also, in assigning subjects to the experimental and control groups in Study 2, an effort was made to keep the groups equivalent for actual weight and percent overweight. In Study 1, however, subjects were randomly distributed.

ARE WALKING (instead of PEOPLE WALKING) and was printed on two lines instead of one. This change was made so that the control stimulus would be more structurally similar to the experimental stimulus MOMMY AND I ARE ONE, which was printed on two lines.

Procedure. Four graduate psychology students not associated with the first study served as interventionists. Three were female and one was male, and each saw an equal percent of subjects in the experimental and control groups. As in the first study, the interventionists were blind to the tachistoscopic condition for each subject.

The procedure was identical to that used in Study 1, with the exception that it was extended from 8 weeks to 12 weeks. Also, a second follow-up weigh-in was conducted 8 weeks after the first. Thus, weights were recorded initially, at 12 weeks (when the behavior program and subliminal stimulation ended), after 16 weeks (follow-up 1), and after 24 weeks (follow-up 2). Additionally, to determine whether weight loss would be accompanied by symptom substitution, the subjects also were given a symptom rating scale to fill out at the same four points in time. This was a variant of the Symptom Check List (90 items) (Derogatis, Lipman, Rickels, Uhlenhuths, & Covi, 1974) in which they were asked to indicate on a 5-point scale the degree to which each of 49 psychiatric symptoms were present.⁷

Finally, at the time of the second follow-up, information was elicited from the subjects about the extent to which they had made use of the two aspects of the therapy program during the prior 24 weeks, *outside of the treatment sessions*. That is, they were asked to rate on a 10-point scale, ranging from "not at all" (1) to "extremely frequently" (10), their average weekly use of (a) the behavior modification techniques and (b) the practice of forming mental images of the flashes of light when they were trying to refrain from eating.

Results

Tables 3 and 4 present the data on the actual weight and percent overweight for Study 2. Analyses of covariance revealed that these results closely parallel those from Study 1. At the end of treatment, the differences between the experimental and control groups were not significant on either measure, $F(1, 23) = 2.42$, $p = .130$, and $F(1, 23) = 1.86$, $p = .183$, for actual weight and percent overweight, respectively. However, at the first follow-up 4 weeks later (or 16 weeks after treatment began), both measures attained significance, $F(1, 23) = 5.03$, $p = .033$, and $F(1, 23) = 4.65$, $p = .039$, respectively. At the second follow-up 8 weeks after the first (or 24 weeks after the program began), the

Table 3

Means and Standard Deviations of Weight in Pounds for Study 2

Time	Experimental group		Control group	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Pretreatment	168.6	25.7	171.8	33.5
Posttreatment (12th week)	156.8	25.9	164.7	32.5
Follow-up 1 (16th week)	154.6	26.7	165.1	26.1
Follow-up 2 (24th week)	153.5	27.5	165.9	28.2

Note. 1 pound = .4536 kg.

difference between the two groups was again significant on both measures, with the means even further apart than they were at the time of the first follow-up, $F(1, 23) = 7.46$, $p = .011$, and $F(1, 23) = 7.40$, $p = .012$, for actual weights and percent overweight, respectively. As in Study 1, the significant differences at the time of follow-up were due to the experimental group subjects', on the average, continuing to lose weight after the behavior treatment sessions ended, whereas the control group gained back some of the weight that had been lost.

On the symptom rating scale, both groups showed a significant reduction in pathology reported from the initial assessment to the 16-week assessment ($t = 2.84$, $p = .015$, and $t = 3.03$, $p = .011$ for the experimental and

⁷ However, sufficient data were available for comparing the experimental and control subjects only initially and at Follow-up 1. At these times the subjects filled out the rating scale in our laboratory, whereas for the postassessment and for Follow-up 2, many subjects took the scales home with them but never returned them. It also should be noted that at Follow-up 2, five subjects (two experimental and three control) could not return to our laboratory for a weigh-in either because they were out of town or because their job situations would not allow it. We contacted them by telephone and asked them to weigh themselves elsewhere and then phoned them a day later for their weights. Although we cannot be certain that the weights they reported would coincide with what their weights would have registered on our own scale, they followed the same pattern as the subjects in their stimulus groups who came in for the weigh-in.

Table 4
Means and Standard Deviations of Percent Overweight for Study 2

Time	Experimental group		Control group	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Pretreatment	37.9	26.7	36.3	26.6
Posttreatment (12th week)	28.5	26.5	30.3	26.8
Follow-up 1 (16th week)	26.5	27.9	30.5	26.3
Follow-up 2 (24th week)	25.9	28.3	31.6	25.0

control groups, respectively). The difference in symptom reduction between the two groups was nonsignificant. (With the premeasure covaried out, $F = .024$, $p = .99$.)

Finally, for their ratings (on a 10-point scale) of the degree to which they used the two aspects of the therapy program outside of the treatment sessions, the differences between the two groups were also negligible. With regard to the behavior modification techniques, the average ratings for the experimental and control groups were, respectively, 7.9 and 7.1, and for the use of the mental images of the flashes, 2.8 and 3.2 ($t < 1$ in both instances).

Discussion

It seems warranted to conclude that the 4-msec exposure of MOMMY AND I ARE ONE over several weeks aided both groups of experimental subjects in losing weight. It is true that this conclusion is based on the follow-up weights, but this does not detract from its significance, since there is nothing to indicate that anything other than the subliminal stimuli exposed during the treatment program acted differentially on the groups of experimental and control subjects. In both studies the experimental subjects were given the same behavioral therapy program as the controls, and they had the same number of tachistoscopic exposures by interventionists who were blind to the stimulus messages. In the second study, there were self-report data that indicated that there was no difference in the frequency with which the two groups,

outside of the treatment sessions, used the techniques that they had been taught. Further, since the subjects themselves did not know what stimuli they were receiving, demand characteristics cannot be implicated to account for the significant differences at follow-up. Only the actual content of the stimuli that were subliminally exposed was different for the experimental and control subjects.

The findings from the two studies reported here should be considered in combination with analogous results from investigations with (insect) phobics (Silverman et al., 1974), schizophrenics (Silverman et al., 1975), and undergraduate college students (Parker, 1977) cited earlier. In toto, the results indicate that the subliminal stimulation of the fantasy MOMMY AND I ARE ONE not only can lead to improved adaptation within a laboratory session, as earlier studies of subliminal psychodynamic activation have demonstrated (summarized in Silverman, 1976), but when combined with an intervention that is effective in its own right and when given over a prolonged period of time, this effectiveness is enhanced outside the laboratory. Thus, this intervention can be viewed as having practical utility.

As far as the intervention's utility as an aid in weight control is concerned, it should be noted in the tables that the significant differences that were found were a function of the fact that while the control groups, on the average, regained weight during the follow-up periods, both experimental groups showed further weight loss; and in Study 2, even more weight was lost during the second follow-up period than during the first. This is noteworthy, since as Hall, Hall, Hanson, and Borden (1974) pointed out, it is unusual for obese patients to continue to lose weight during follow-up periods. An examination of the individual data reveals that although only 32% of the control subjects (in both studies combined) accomplished this, 84% of the experimental subjects continued to lose weight. Thus, this experimental intervention may be able to reverse a trend that typically limits the effectiveness of behavior modification treatment of obesity. On the other hand, it also should be noted that the largest (average) weight loss reported for an experimental

group in this pair of studies (15.1 pounds (6.8 kg) or a 12% reduction in weight in Study 2 over a 24-week period) must be viewed as modest and not beyond the range of what has been reported by other (relatively) successful intervention programs for weight control. Thus, if subliminal symbiotic stimulation is to be of substantive value in the treatment of obesity, ways should be sought to increase its potency.

A comparison of the data from Study 1 with that from Study 2 underscores the fact that what led to the differential behavior of the experimental and control subjects during the follow-up periods was not the additional time that elapsed but their response to the nonavailability of treatment. By extending the program an additional 4 weeks in Study 2, the time between the preweights and postweights (12 weeks) was identical to the time between the preweights and follow-up weights in Study 1. Yet even though this extra 4 weeks led to increased weight loss for *both* groups (note in the tables the approximately 30% greater posttreatment weight loss in Study 2 than in Study 1), it did not differentially affect the experimental and control subjects. Thus in Study 2, as in Study 1, the difference between the experimental and control groups did not reach significance until the subjects were on their own for 4 weeks. Apparently, the crucial benefits of the symbiotic stimulation was in allowing the subjects to retain their ability to diminish food intake *in the absence of weekly treatment contact*. Or in psychoanalytic terminology, the presumed activation of the MOMMY AND I ARE ONE fantasy may have allowed the subjects to better internalize the techniques that they had been taught. Whether this was because the therapist was unconsciously equated with "mommy" (whom they now felt more at one with) or for some other reason remains a matter for further study.

A number of other questions also remains to be addressed. First, it would be important to determine from continued follow-ups the duration of the "boost" that the subliminal symbiotic stimulation gives to behavior modification methods in controlling overeating.

Second, there is the question of whether the

diminished overeating brought about by the symbiotic stimulation is at the expense of personality change that could be viewed as maladaptive. In Study 2, the fact that there was no difference in symptom reduction between the experimental and control groups and that, in fact, both groups reported significantly *fewer* symptoms at 16 weeks than they did initially strongly argues against the operation of symptom substitution. However, as one of us has argued elsewhere (Silverman, 1974) in evaluating whether symptom reduction is "gained at a price" as a result of *any* kind of intervention, more than symptom substitution must be evaluated. From what can be observed clinically, the disappearance of symptoms is sometimes accompanied by the emergence of maladaptive behavior that is not experienced as a "symptom." In evaluating any kind of therapeutic intervention, one should investigate whether, and if so to what degree, asymptomatic as well as symptomatic negative personality changes occur when symptoms remit, a practice that we plan to follow in investigations using subliminal symbiotic stimulation.

Third, there is the question of what the precise aspects are of the MOMMY AND I ARE ONE stimulus that account for its effectiveness in weight control. In work with other subject groups, data from several studies have indicated that in order for this stimulus to be ameliorative (when subliminally presented), it must contain a reference to "oneness," but the inclusion of mommy as the person with whom the oneness is achieved is not essential.⁸

⁸ With regard to the results on "oneness," there were two relevant studies. Kaplan (1976) found that while MOMMY AND I ARE ONE reduced pathology in schizophrenics, other reassuring messages involving mommy (e.g., MOMMY FEEDS ME WELL and MOMMY IS ALWAYS WITH ME) did not have this effect. Bronstein (1976) investigated the effectiveness of other "internalization" of mother messages (e.g., MOMMY IS INSIDE ME and MOMMY AND I ARE ALIKE) as well as MOMMY AND I ARE ONE and found that only the latter reduced pathology for schizophrenics. Even though the reference to oneness thus seems to be essential for an adaptation-enhancing effect, data from two other studies indicate that the fantasy of oneness does not have to involve mommy. Kaye (1975) found that the stimulus MY GIRL AND I ARE

Whether the same applies to obese individuals is something we are currently investigating.

There is also an ethical question that can be raised about subjecting people to subliminal stimulation designed to affect their behavior. This issue was dealt with in the current studies by debriefing the subjects at the experiment's conclusion as to the content of their subliminal message and providing them with the opportunity to talk about their reactions. More recently, we have been informing subjects from the beginning that they will be receiving subliminal messages and have found that this does not prevent the MOMMY AND I ARE ONE message from producing its ameliorative effect (Parker, 1977). A further step toward complete openness with subjects that currently is being explored is telling the subjects beforehand which subliminal messages are being used, without telling them which particular message he or she is receiving.

Finally, a question can be raised about what mediates the effectiveness of the MOMMY AND I ARE ONE message as an aid in weight control. Does it act synergistically with the behavior modification techniques, making it easier for subjects to learn and use the latter? Or does it act in a more direct manner by strengthening behavior controls, reducing anxiety, or even by diminishing unconscious symbiotic longings? To make this determination, measures will have to be obtained of the variables just cited in each treatment session, a research strategy that we plan to pursue.

Reference Note

1. Silverman, L. H. *The unconscious symbiotic fantasy as a ubiquitous therapeutic agent*. Paper presented at the meeting of the American Psychoanalytic Association, Baltimore, Maryland, May 1976.

References

- Bronstein, A. An experimental study of internalization processes in schizophrenic men. (Doctoral dissertation, Yeshiva University, 1976). *Dissertation Abstracts International*, 1976, 37, 4665B. (University Microfilms No. 77-5001)
- Bruch, H. *Eating disorders*. New York: Basic Books, 1973.
- Bychowski, G. On neurotic obesity. *Psychoanalytic Review*, 1950, 27, 301-319.
- Condon, T. Systematic desensitization: An evaluation of a psychoanalytic model of its effectiveness. (Doctoral dissertation, University of Connecticut, 1976). *Dissertation Abstracts International*, 1976, 38, 349B. (University Microfilms No. 77-14, 454)
- Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., & Covi, L. The Hopkins Symptom Checklist (HSCL): A measure of primary symptom dimensions. In P. Pinchot (Ed.), *Psychological measurements in psychopharmacology: Modern problems in pharmacopsychiatry* (Vol. 7). Basel, Switzerland: Karger, 1974.
- Dixon, N. F. *Subliminal perception: The nature of a controversy*. London: McGraw-Hill, 1971.
- Hall, S. M., Hall, R. G., Hanson, R. W., & Borden, B. L. Permanence of two self-managed treatments of overweight in university and community populations. *Journal of Consulting and Clinical Psychology*, 1974, 42, 781-786.
- Kaplan, R. The symbiotic fantasy as a therapeutic agent: An experimental comparison of the effects of three symbiotic elements on manifest pathology in schizophrenics. (Doctoral dissertation, New York University, 1976). *Dissertation Abstracts International*, 1976, 37, 1437B. (University Microfilms No. 76-19, 512)
- Kaye, M. The therapeutic value of three merging stimuli for male schizophrenics. (Doctoral dissertation, Yeshiva University, 1975). *Dissertation Abstracts International*, 1975, 36, 1437B. (University Microfilms No. 75-20, 571)
- Leiter, E. A study of the effects of subliminal activation of merging fantasies in differentiated and non-differentiated schizophrenics. (Doctoral dissertation, New York University, 1973). *Dissertation Abstracts International*, 1973, 34, 4022B. (University Microfilms No. 74-01919)
- Mahler, M., Pine, F., & Bergman, A. *The psychological birth of the human infant*. New York: Basic Books, 1975.
- Martin, A. The effect of subliminal stimulation of symbiotic fantasies on weight loss in obese women receiving behavioral treatment. (Doctoral dissertation, New York University, 1975). *Dissertation Abstracts International*, 1975, 36, 3054B. (University Microfilms No. 75-28568)
- Parker, K. *The effects of the subliminal stimulation of symbiotic fantasies on the school performance of college students*. Unpublished doctoral dissertation, New York University, 1977.
- Sackeim, H. Self-deception: Motivational determinants of the non-awareness of cognition. Unpub-

ONE also produced an amelioration of symptoms in male schizophrenics; and Parker (1977) in the earlier cited study of the effects of subliminal stimulation as a learning aid for college students, found that the stimulus MY PROF AND I ARE ONE, as well as MOMMY AND I ARE ONE, led to significantly higher grades than a neutral control stimulus.

- ished doctoral dissertation, University of Pennsylvania, 1977.
- Silverman, L. H. Drive stimulation and psychopathology: On the conditions under which drive-related external events evoke pathological reactions. In R. R. Holt & E. Peterfreund (Eds.), *Psychoanalysis and contemporary science* (Vol. 1). New York: Macmillan, 1972.
- Silverman, L. H. Some psychoanalytic considerations of non-psychoanalytic therapies: On the possibility of integrating treatment approaches and related issues. *Psychotherapy: Theory, Research and Practice*, 1974, 11, 298-306.
- Silverman, L. H. Psychoanalytic theory "The reports of my death are greatly exaggerated." *American Psychologist*, 1976, 31, 621-637.
- Silverman, L. H., & Candell, P. On the relationship between aggressive activation, symbiotic merging, intactness of body boundaries and manifest pathology in schizophrenics. *Journal of Nervous and Mental Disease*, 1970, 150, 387-399.
- Silverman, L. H., Frank, S. G., & Dachinger, P. A psychoanalytic reinterpretation of the effectiveness of systematic desensitization: Experimental data bearing on the role of merging fantasies. *Journal of Abnormal Psychology*, 1974, 83, 131-138.
- Silverman, L. H., Kwawer, J. S., Wolitzky, C., & Coron, M. An experimental study of aspects of the psychoanalytic theory of male homosexuality. *Journal of Abnormal Psychology*, 1973, 82, 178-188.
- Silverman, L. H., Levinson, P., Mendelsohn, E., Ungaro, R., & Bronstein, A. A clinical application of subliminal psychodynamic activation: On the stimulation of symbiotic fantasies as an adjunct in the treatment of hospitalized schizophrenics. *Journal of Nervous and Mental Disease*, 1975, 161, 379-392.
- Silverman, L. H., Spiro, R. H., Weissberg, J. S., & Candell, P. The effects of aggressive activation and the need to merge on pathological thinking in schizophrenia. *Journal of Nervous and Mental Disease*, 1969, 148, 39-51.
- Spiro, T. The effects of laboratory stimulation of symbiotic fantasies and bodily self-awareness of relatively differentiated and non-differentiated schizophrenics. (Doctoral dissertation, New York University, 1975). *Dissertation Abstracts International*, 1975, 36, 5818B. (University Microfilms No. 76-10, 227)
- U.S. Department of Health, Education, and Welfare. *Obesity and health: A sourcebook of current information for the professional health personnel*. Arlington, Va.: U.S. Public Health Service, 1967.
- Wollersheim, J. P. Effectiveness of group therapy based upon learning principles in the treatment of obese women. *Journal of Abnormal Psychology*, 1970, 76, 462-474.

Received February 1, 1977 ■

Specific and Nonspecific Factors in the Effectiveness of a Behavioral Approach to the Treatment of Marital Discord

Neil S. Jacobson
University of Iowa

This study compared two behavioral treatments for marital discord with a non-specific control and a waiting-list control. The behavioral treatments combined training in problem-solving skills with training in contingency management procedures, differing only with respect to the contracting form: One group learned to form good faith contracts, and the other, quid pro quo contracts. Thirty-two couples were randomly assigned to one of these treatment conditions and one of three therapists. Improvement was assessed by two observational measures and by two self-report questionnaires. On all measures, both behavioral groups improved significantly more than waiting-list couples. On three of the four measures, behavioral couples improved significantly more than nonspecific couples. The two behavioral groups did not differ from one another on any of the measures.

Although an increasing number of reports have attested to the effectiveness of using social-learning principles in treating marital problems, most of these reports consist of uncontrolled case studies (cf. Jacobson & Martin, 1976). Those few controlled studies attempting to evaluate the effectiveness of such approaches have often been limited

either by equivocal findings (e.g., Harrell & Guernsey, 1976), methodological inadequacies (Tsoi-Hoshmand, 1976), or the analogue nature of the study (Margolin, Note 1).

Despite the lack of definitive evidence in favor of a behavioral approach to marital therapy, the proliferation of suggestive findings warrants cautious optimism. Particularly promising is the combination of training in communication and problem-solving skills with instructions in contingency management procedures, especially contingency contracting (Weiss, Birchler, & Vincent, 1974). A series of impressive case reports (Weiss, Hops, & Patterson, 1973), in addition to Margolin's (Note 1) study, attest to the effectiveness of this combined treatment package.

Jacobson (1977a) evaluated the effectiveness of such a treatment program, comparing it with a minimal treatment, waiting-list control group. On both observational and self-report measures, the couples receiving the behavioral treatment package improved significantly more than did control couples. Within-subject analyses, using data collected by spouses in the home, tended to corroborate the group comparisons. A 1-year follow-up suggested that these positive changes had been maintained.

This manuscript was based on the author's doctoral dissertation, conducted as partial fulfillment of the requirements for a PhD degree. Portions of this article were presented at the meeting of the Association for the Advancement of Behavior Therapy, New York, December 1976. The research was supported by National Institute of Mental Health Research Fellowship 1F31MH05287-01 and by a grant from the Oregon Research Institute.

The author wishes to thank Barclay Martin, the chairperson of his dissertation committee, along with other members of the committee, Judith Flaxman, David Eckerman, Robert Johnson, William Eichman, and William Anderson, for their valuable advice and support. Special thanks are extended to W. Bruce Willingham, who performed invaluable service as chief research assistant; other research assistants to whom the author is indebted are Casey Jacob, Sharon Lorch, Ann Berry, Kurt Nelson, Tim Rodell, and Randy Bernard.

Requests for reprints should be sent to Neil S. Jacobson, Department of Psychology, University of Iowa, Iowa City, Iowa 52242.

Only limited conclusions can be drawn from this initial inquiry. In addition to the small sample size and the inclusion of only one therapist, it is unclear which factors accounted for the effectiveness of the experimental treatment.

The present study served three purposes, two of which followed directly from the limitations of the prior study (Jacobson, 1977a). First, the prior study was replicated, using three therapists instead of one. Second, a group was included that was designed to control for nonspecific factors in the experimental treatment. The group was carefully constructed to control for as many possible competing explanations for change as was feasible, using some of the guidelines suggested by Jacobson and Baucom (1977).

The third purpose of the study was to shed light on a controversy regarding the type of contingency contracting that is most efficacious in the treatment of marital discord. Some (cf. Jacobson & Martin, 1976) have advocated a contracting procedure that Weiss et al. (1974) have referred to as "quid pro quo" contracting. Both partners agree to make a change that the other has deemed desirable; one partner's change serves as a reinforcer for the other's change. Weiss et al. (1974) cautioned against this form of contracting in severely distressed relationships: They argued that the contingent relationship between each spouse's behavior change agreements creates a "who goes first" problem; neither partner is likely to change under such circumstances, given the degree of mistrust in severely distressed relationships. Another potential hazard of quid pro quo contracting is that a failure to change by one spouse sanctions the abdication of the other's contractual responsibilities.

Weiss et al. (1974) advocated an alternative form of contracting, referred to as "parallel" or "good faith" contracting. Here each spouse independently agrees to initiate a change in behavior desired by their partner; rather than the changes being cross-linked to one another, each change is independently reinforced and/or punished.

In Jacobson's initial study (1977a), good faith contracts were used. However, there is no empirical evidence as yet to recommend

the good faith option. Good faith contracts carry with them a distinct disadvantage: They are less efficient, requiring the pinpointing of effective reinforcers and punishers outside the set of target problems defined by the partners. The specification of such consequences can be a cumbersome task. Thus, the present study sought to compare the effectiveness of quid pro quo and good faith contracting procedures.

Couples seeking assistance for relationship problems were randomly assigned to one of three therapists and one of four treatment conditions: a behavioral treatment program using good faith contracts (GF group), a behavioral treatment program using quid pro quo contracts (QPQ group), a nonspecific treatment condition (NS), and a waiting-list control group (WL).

Method

Subjects

Client couples were solicited through three sources: (a) advertisements placed in local newspapers; (b) public service announcements on local radio stations; and (c) referrals from mental health agencies and individual clinicians. The vast majority of couples ($n = 32$) who eventually participated in the study responded to one of the advertisements. The advertisement offered a new experimental treatment for couples experiencing marital problems. Couples were interviewed provided that they had completed the Marital Precounseling Inventory (Stuart & Stuart, 1973), which was mailed to them before the scheduled time of the initial interview.

Therapists

Three therapists treated couples in the study. Two were advanced graduate students in clinical psychology; the other was a master's level social worker. All three had between 2 and 3 years of clinical experience.

I was one of the advanced graduate student therapists. The other two received about 20 hours of training each from me, along with 1-1½ hours of weekly supervision once the study began.

During the first 4 months of the study, couples were randomly assigned either to Therapist 1 or Therapist 2. During the second 4 months of the study, couples were randomly assigned to either Therapist 1 or Therapist 3.

Measures

Marital Interaction Coding System (MICS; Hops, Wills, Patterson, & Weiss, 1971). This was the same behavioral rating system used and described by

Jacobson (1977a). A modified version was used in the present study. Instead of using all 30 behavioral categories, couples' behavior was coded according to 1 of 12 categories; only verbal behavior was coded.

The behavior coded by the MICS consisted of interaction by couples in problem-solving situations before and after therapy. On each occasion, couples were given 5-10 minutes to solve a hypothetical marital problem from the Inventory of Marital Conflicts (Olson & Ryder, 1970) and a "minor" problem in their relationship, which couples chose themselves. These discussions were videotaped and later coded by trained raters who were kept blind to experimental conditions. Reliability was assessed by dividing the frequency with which the two raters agreed that a particular category should be coded by the sum of agreements and disagreements, that is, the number of times a category was marked by one rater but not by the other. This quotient was then multiplied by 100. Both raters coded each discussion for each couple in the study, so that reliability could be continuously assessed. The reliability quotients for particular problem-solving discussions ranged from .56 to 1.00, with an average reliability quotient of .84.

For the purposes of analyses, the 12 categories were further collapsed into 3: positive behavior, negative behavior, and neutral behavior. Positive and negative behavior, recorded in terms of rate per minute, were the two dependent variables derived from the MICS. Scores for husbands and wives were combined so that each couple was analyzed as a unit. Similarly, for a given session, the hypothetical and "real" problems discussed were combined for analysis.

Marital Adjustment Scale (MAS; Locke & Wallace, 1959). This traditional self-report index of marital adjustment was administered to spouses before and after therapy. The scale provides an overall rating of marital satisfaction for each spouse, with a score of 100 (or 200 for a couple) as the cutoff between a satisfying and unsatisfying relationship. For the purposes of analysis, the combined score for the couple was used. This test was administered to couples before and at the conclusion of therapy.

Marital Happiness Scale (MHS). This rating scale is a subscale of Stuart and Stuart's (1973) Marital Pre-Counseling Inventory; couples are required to rate their degree of happiness in regard to 12 general categories of marital life. Since not all 12 items were applicable to all couples, a score for a given spouse was derived by dividing the sum total of the rating for each item by the total number of applicable items, yielding an average rating per item. Then spouse scores were combined, and each couple was analyzed as a unit. Lower scores indicated greater "happiness": The minimum score was 1.00, whereas a maximum score (suggesting maximum "unhappiness") was 5.00.

Manipulation Checks

Posttherapy session evaluation. After each therapy session, all spouses rated both the therapy session and

the therapist on a variety of dimensions, using a questionnaire adapted from Lazarus (1971). Clients' perceptions of the therapist were used as a means of evaluating the degree of stylistic and procedural similarity across therapists. In addition, certain items were used as indirect indices of the credibility of the various treatment conditions.

Descriptions of treatment procedures rated by undergraduates. A group of undergraduate students were randomly presented one of two written treatment descriptions: One described the behavioral treatment condition (without specifying the type of contracting form used), and the other described the non-specific treatment condition. Subjects then completed a brief questionnaire designed to assess the credibility of the procedure as perceived by each subject.

Procedure

At the conclusion of an initial interview, in which a brief history was taken, a set of problems was defined, pretesting was administered (MICS, MAS), and couples were randomly assigned to one of four treatment conditions: GF, QPQ, NS, or WL. WL couples were told that simply experiencing the evaluation and the initial interview had been helpful to some couples, and that they should wait for 8 weeks and then return to decide whether or not they wanted further treatment. Such treatment was offered to couples on their return for posttesting.

For the couples in the other three conditions, a brief description of the research aspects of the program was offered. Procedures subsequently varied depending on the condition to which the couple had been assigned. All treatment groups met weekly for 1-1½ hours, and there were eight treatment sessions in addition to the initial interview.

NS group. Couples in this condition received a treatment devoid of specific instructions in problem-solving and communication skills and without contingency contracting procedures. In all other respects, this condition was designed to duplicate the two behavioral treatment conditions. The condition was roughly equivalent to the behavioral treatment program on the following stylistic and procedural variables:

1. Attention.
2. Expectancies of therapeutic gain. Optimistic and confident statements were identical to those directed toward GF and QPQ couples. The degree to which therapists succeeded in this endeavor was determined by clients' ratings of their therapist following each therapy session. Couples were asked to rate the degree to which their therapists seemed "optimistic" and "confident."
3. Credibility of treatment procedures. Every effort was made to design a treatment condition that would be perceived as credible. To obtain comparative credibility ratings of the various treatment conditions that would not be confounded with outcome, brief descriptions of each treatment procedure were rated as to their credibility by undergraduates.

4. Therapist activity level. In the experimental treatment conditions, the therapist spoke frequently. Similarly, in the NS condition therapists were asked to attempt to intervene as often as they did in GF and QPQ conditions.

5. Directiveness-nondirectiveness. One frequent source of confounding in comparisons between experimental and nonspecific treatment conditions involves the extent of therapist directiveness (cf. Jacobson & Baucom, 1977). Often, therapists in nonspecific conditions are less directive than in corresponding experimental conditions; this difference might be responsible for any discrepancies in treatment effectiveness observed between two groups. Since in the experimental treatment groups the therapy sessions were structured and the therapist actively initiated topics of conversation, such conditions were maintained for NS couples. Client's perceptions of therapy sessions were used to evaluate the therapists' level of directiveness; in posttherapy session questionnaires, clients were asked, for example, "To what extent did the therapist initiate topics of conversation?"

In addition, therapy sessions in the NS condition were structured. In fact, procedurally, the formats of the various conditions were virtually identical. Couples in all treatment conditions were asked to record their change agreements in writing. Also, NS couples engaged in homework assignments between treatment sessions.

6. Opportunity to discuss marital problems.

7. Presentation of a "rationale."

At the conclusion of the initial interview, NS couples were asked to record all "affectionate" acts initiated by the other and to designate each as either "pleasing" or "displeasing." This assignment was ongoing until the end of therapy.

During the initial treatment session, couples were presented with an explanation of the treatment package. The program was presented as an attempt to maximize the effectiveness of many behavior change strategies, each of which by itself has proven somewhat effective in bringing about desirable changes.

The treatment sessions were structured such that problems specified during the initial interview were discussed. Each problem was discussed by the couple until some agreement was reached. The therapist participated actively in such discussions, but at no time did he (she) suggest strategies for solving a particular problem. Nor did he (she) present couples with specific behavioral feedback regarding their performance.

Rather, the therapist's responses were of four types: first, he (she) asked factual questions; second, he (she) restated clients' verbalizations with an emphasis on the apparent "affect" underlying them; third, he (she) made interpretative comments on the basis of the interaction process between the spouses; fourth, the therapist could emit self-disclosing remarks either by offering them a personal reaction to some event in therapy or by offering examples from his (her) own life.

The structure of treatment sessions was identical to

that of the behavioral conditions. At the conclusion of the 8-week program, all couples were offered the opportunity to receive further therapeutic assistance; referrals were arranged by therapists in the study if requests for further therapy were made.

GF group. The treatment program that these couples received replicated the experimental treatment used in the prior study (Jacobson, 1977a) and is described in detail by Jacobson (1977a). Briefly, the first session began with an explanation of the treatment program, including a theoretical rationale. For the remainder of that session and Sessions 2 and 3, couples were taught communication and problem-solving skills, according to guidelines described by Jacobson (1977b, 1977c). Beginning with Session 4, good faith contracting strategies were introduced; for the duration of the treatment program, couples were taught to end their problem-solving sessions with a written agreement specified in the form of a good faith contract. The last treatment session served as a posttest, during which the MAS, MHS, and problem-solving discussions (coded using MICS) were repeated.

Throughout the treatment procedures, couples were assigned problem-solving practice sessions at home between therapy sessions. They kept a notebook to record the details of each practice session.

Skills such as problem solving and good faith contracting were taught using coaching or modeling, and behavior rehearsal (cf. Jacobson, 1977c).

QPQ group. Couples in this condition received treatment identical to that received by GF couples, with one exception: Change agreements taught to couples in this group took the form of quid pro quo exchanges (cf. Weiss et al., 1974). From Session 4 until the end of the treatment program, couples were taught to make exchanges, whereby the husband would make a change desired by the wife in return for a change by the wife desired by the husband. The specified reinforcer in every instance was the change agreed to by the other spouse.

Results

Data Analysis

The primary statistical technique used was multivariate analysis of covariance, with pretest scores on each of the criterion variables used as covariates. Only couples who completed all treatment sessions and were available for posttesting were included in the analysis for treatment effects. Therapist effects as well as Therapist \times Treatment interactions were analyzed with the WL condition excluded. Dropouts were included in this analysis; if such couples were unavailable for posttesting, no change was assumed and posttest scores were assigned that were identical to

pretest scores. Given the small number of couples in each treatment cell, to exclude dropouts would have falsified actual therapist performance, since these were the couples who responded least favorably to treatment. The inclusion of dropouts added two couples to the total sample, adding one GF couple to Therapist 3 and one NS couple to Therapist 2.

Planned comparisons were used for the three individual comparisons of greatest interest (see below), using nonorthogonal contrasts with the alpha level set at .02. Planned comparisons were analyzed using analysis of covariance, with pretest scores serving as covariates. Post hoc comparisons were analyzed using Tukey's honestly significant difference test. For these comparisons, pre-post difference scores were used. Other analyses are described below.

Pretreatment Characteristics of Couples

Demographic characteristics. Couples in the various treatment conditions were compared on age of husband, age of wife, duration of marriage, number of children, and education level for both husband and wife. A multivariate analysis of variance failed to un-

cover differences between treatment conditions on these various measures.

The age of husbands averaged 32.9 years; wives' age averaged 31.8 years. Couples had been married for an average duration of 7.9 years. The number of children averaged 1.32 per couple.

Pretreatment differences on criterion measures. Table 1 shows pretreatment means for all treatment conditions on the four criterion measures used in the study. The overall multivariate test comparing the groups on these measures was nonsignificant.

Therapist Effects

Therapist effects were analyzed by conducting a multivariate analysis of covariance with pretest scores on all dependent measures serving as covariates. WL couples were excluded from this analysis. The test for therapist differences was nonsignificant, $F(8, 20) = .971, p > .50$.

The overall multivariate test for the Therapist \times Treatment interaction was similarly nonsignificant, $F(16, 31) = .41, p > .75$.

Thus, there was no evidence of differential therapist performance. Nor was there any in-

Table 1
Pretreatment and Posttreatment Scores on the Four Dependent Measures for the Four Treatment Groups

Group	n	Negative behavior ^a		Positive behavior ^a		MAS		MHS	
		Pre	Post	Pre	Post	Pre	Post	Pre	Post
GF	8								
M		6.60	3.27	3.08	4.99				
SD		2.20	1.75	1.00	1.50	173.50	214.13	2.99	2.02
QPQ	9					29.67	31.56	.07	.50
M		5.82	3.45	3.16	5.09				
SD		2.39	1.70	1.12	2.32	158.56	205.56	2.99	2.31
NS	7					20.73	33.37	.49	.59
M		6.27	5.64	3.73	3.23				
SD		2.44	2.42	2.49	1.62	167.57	173.56	3.14	2.53
WL	6					44.27	50.33	.68	.90
M		6.18	7.81	3.88	2.23				
SD		1.46	2.11	1.69	1.28	140.68	123.5	3.40	3.47
						23.01	25.34	.27	.50

Note. Couples were analyzed as a single unit; each mean score represents the combination of a husband's score and his wife's score. GF = good faith; QPQ = quid pro quo; NS = nonspecific treatment; WL = waiting list; MAS = Marital Adjustment Scale; MHS = Marital Happiness Scale.

^a Frequency of behavior emitted (rate per minute).

dication of change in relative therapist performance as a function of treatment condition. Subsequent analyses ignored the therapist factor.

Treatment Effects

A multivariate analysis of covariance was conducted to determine whether an overall treatment effect existed; in this analysis, pretest scores on the four dependent measures were used as covariates. The analysis indicated that a treatment effect did exist, $F(12, 50.56) = 3.36$, $p < .001$. Subsequent analyses specifying the nature of the treatment effect are presented below for each of the criterion measures separately.

Negative behavior. Table 2 indicates the significance tests for each of the four treatment groups in regard to the null hypothesis that the pretest-posttest difference scores were equal to zero for negative behavior. As Figure 1 indicates, the two behavioral treatment groups (GF and QPQ) as well as the

Table 2

Significance Tests Regarding Pretest-Posttest Difference Scores for Each Treatment Condition on Each of the Four Criterion Measures

Group and measure	MS	F
GF		
Negative behavior	88.511	22.507***
Positive behavior	25.49	6.983*
MAS	13,203.145	23.938***
MHS	7.644	55.806***
QPQ		
Negative behavior	50.505	12.842***
Positive behavior	34.496	9.451**
MAS	20,258.711	36.730***
MHS	30.881	30.881***
NS		
Negative behavior	.167	.042
Positive behavior	1.760	.482
MAS	206.281	.374
MHS	18.310	18.310***
WL		
Negative behavior	15.974	4.062
Positive behavior	16.236	4.448
MAS	2,128.173	3.858
MHS	.028	.205

Note. MAS = Marital Adjustment Scale; MHS = Marital Happiness Scale. For all tests, $df = 1, 26$.

* $p < .01$.

** $p < .005$.

*** $p < .001$.

NS group manifested, on the average, fewer negative behaviors at posttest than during pretest; in contrast, the WL group deteriorated. For both the GF and QPQ groups, the changes were highly significant, whereas for the NS group the changes were nonsignificant. The WL couples' trend toward deterioration bordered on significance. Thus, the only groups that changed significantly in the desired direction, in terms of the frequency of negative behavior, were the GF and QPQ groups.

Planned comparisons of greatest interest were formed. As Table 3 indicates, the two behavioral groups combined were significantly more effective than the WL condition, $F(1, 25) = 31.17$, $p < .001$. Also, the two behavioral groups were significantly more effective than the NS group, $F(1, 25) = 8.80$, $p < .007$. However, there were no significant differences between the two behavioral groups, $F(1, 25) = .49$, $p < .45$.

Post hoc analyses were used for the remain-

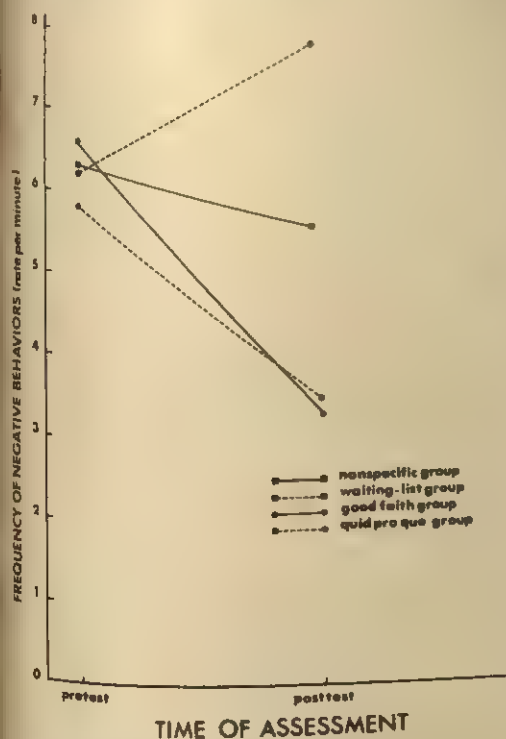


Figure 1. Changes in negative behavior for each of the four treatment conditions.

ing paired comparisons.¹ For each comparison, a q statistic was computed using Tukey's honestly significant difference test. Both GF couples and QPQ couples improved significantly more than did WL couples ($p < .01$). However, only GF couples improved significantly more than did NS couples ($p < .01$); although QPQ couples also changed more than did NS couples, the differences did not reach statistical significance. Differences between NS and WL conditions were nonsignificant.

Positive behavior. Figure 2 depicts increases in positive behavior from pretest to posttest for each of the treatment conditions. Again, both behavioral groups changed substantially in the desired direction; this time trends toward deterioration occurred in both the NS and WL conditions. As Table 2 indicates, the changes were statistically significant for both the GF group ($p < .01$) and the QPQ group ($p < .005$).

Table 3 lists the planned comparisons for this measure. The analysis of covariance revealed that the behavioral groups were significantly more effective in increasing positive behavior than either the WL group ($p <$

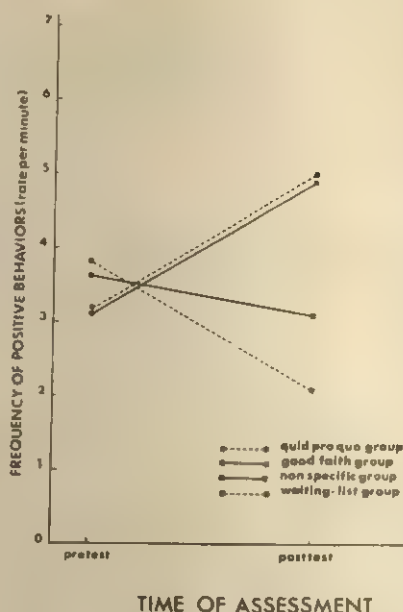


Figure 2. Changes in positive behavior for each of the four treatment conditions.

$p < .001$) or the NS group ($p < .01$). Using post hoc comparisons, the GF group alone was significantly more effective than the WL group ($p < .01$) and the NS group ($p < .05$). Similarly, the QPQ group was more effective than either the WL ($p < .01$) or the NS group ($p < .05$). Neither the two behavioral groups ($p > .90$) nor the two control groups ($p > .05$) differed significantly from one another.

MAS. Figure 3 shows that on this self-report measure the two behavioral groups again improved substantially from pretest to posttest. With husbands' and wives' scores combined, a score of 200 is generally considered to be the cutoff for a normal degree of marital adjustment. All four of the group means were well below this figure prior to therapy. Only the two behavioral groups averaged above 200 subsequent to therapy. The NS group scores remained virtually the same, but again there was a tendency for WL couples to deteriorate. The only changes that were statistically significant, as indicated by Table 2, were those manifested by GF couples ($p < .001$) and QPQ couples ($p < .001$).

Table 3
Planned Comparisons between Treatment Groups
on the Four Criterion Measures

Comparison and measure	MS	F
GF vs. QPQ		
Negative behavior	1.376	.486
Positive behavior	.020	.007
MAS	107.855	.192
MHS	.356	2.635
GF & QPQ vs. WL		
Negative behavior	88.214	31.171****
Positive behavior	41.052	14.131****
MAS	15,752.785	27.982****
MHS	3.443	25.463****
GF & QPQ vs. NS		
Negative behavior	24.910	8.802**
Positive behavior	20.118	6.925*
MAS	7,031.145	12.490***
MHS	.246	1.817

Note. MAS = Marital Adjustment Scale; MHS = Marital Happiness Scale. For all comparisons, $df = 1, 25$.

* $p < .01$.
 ** $p < .007$.
 *** $p < .002$.
 **** $p < .001$.

¹For more details, the reader is urged to write to the author, who will provide a way of obtaining a copy of the complete doctoral dissertation.

As with the two observational measures, on the MAS there were no significant differences between GF and QPQ groups. Taken together, the two behavioral groups improved significantly more than did either the WL group ($p < .001$) or the NS group ($p < .002$). Post hoc comparisons revealed that each of the behavioral groups improved significantly more than did the waiting-list group. (In each instance, $p < .01$.) Similarly, each of the behavioral groups improved more than did the nonspecific group ($p < .05$). There were no significant differences between the WL and NS groups.

MHS. In addition to the GF and QPQ groups, who again showed substantial increases from pretest to posttest on reported happiness, the NS group also showed positive changes on this measure; only the WL failed to report such changes. As Table 2 indicates, in each case the improvement manifested by GF, QPQ, and NS groups was statistically significant.

Similarly, as Table 3 indicates, all three of the treatment groups improved significantly more than did the WL group. However, they did not differ significantly from one another. There was a trend favoring the GF group over the QPQ group, and the QPQ group over the NS group, but neither of these trends reached statistical significance.

Follow-up. An MAS form was mailed to each spouse in the GF, QPQ, and NS groups at 1-month, 3-month, and 6-month intervals following the final treatment session. All couples who returned the forms on at least two of the three occasions were included in the follow-up analysis. Of the original sample, only one couple in each of the three treatment conditions was excluded from the analysis. All three of these couples were treated by Therapist 3.

A follow-up score was derived for each couple by averaging the scores on forms returned by each spouse. These averages were then combined so that each couple could be analyzed as a unit.

Difference scores were then computed for each treatment condition by subtracting pretest scores from follow-up scores. An analysis of variance that compared the combined mean

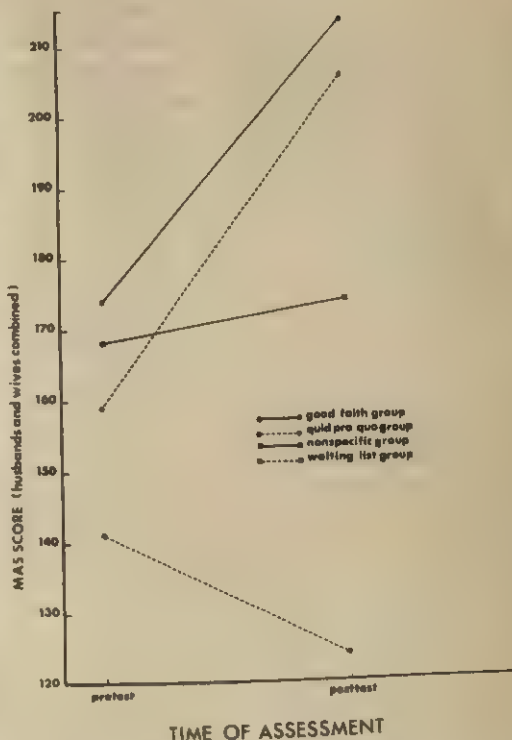


Figure 3. Changes in Marital Adjustment Scale (MAS) for each of the four treatment conditions.

of the two behavioral groups with the mean of the nonspecific group indicated that the groups were significantly different in the direction favoring the behavioral groups, $F(1, 24) = 13.01, p < .002$.

Within-group t tests comparing posttest scores with follow-up scores revealed no significant differences between posttest scores and follow-up scores for any of the three treatment conditions. Thus, the differences favoring the behavioral groups relative to the nonspecific group on the MAS were maintained at the time of the follow-up reports. In addition, group averages on the MAS did not change in any of the three treatment conditions from posttest to follow-up.

Manipulation Checks

Clients' posttherapy session evaluations. After each therapy session, each spouse completed an extensive questionnaire rating the therapist on a variety of dimensions. It was hoped that these ratings would aid in deter-

mining whether (a) therapists performed similarly on relevant stylistic dimensions, (b) behavior was consistent across treatment conditions, (c) the various treatment conditions were equivalent on various nonspecific stylistic and procedural dimensions, and (d) the various treatment conditions seemed equally credible to couples.

A subset of items from the questionnaire was analyzed by a multivariate analysis of variance. The units of analysis were means for each of the 12 items analyzed separately. Husbands' and wives' questionnaires were analyzed separately, and each mean reflected an average rating for a given spouse across therapy sessions. For husbands, the multivariate test was nonsignificant for a treatment effect, $F(24, 8) = .92, p > .50$; a therapist effect, $F(24, 8) = 1.57, p > .25$; and a Therapist \times Treatment interaction, $F(48, 17.45) = 1.27, p > .30$. Similarly, for wives, there was no significant effect for treatment, $F(24, 8) = .95, p > .50$; therapist, $F(24, 8) = 1.58, p > .25$; or Therapist \times Treatment interaction, $F(48, 17.45) = 1.96, p > .05$. Thus, statistical analyses failed to support the notion that clients' posttherapy ratings of their therapists depended on who their therapist was or which treatment condition they were in.

Evaluations of treatment rationales by undergraduate psychology students. After reading a brief description of either the behavioral treatment (with type of contracting unspecified) or the nonspecific treatment, undergraduates answered a seven-item questionnaire designed to assess the apparent credibility of the treatment description. Each of the seven questions was scored separately. Responses for the two groups of subjects were compared using a Hoetelling's T^2 test; the overall test was nonsignificant, $F(7, 121) = .63, p > .70$. On the basis of these responses, it appears that undergraduates found the two treatment rationales equally credible.

Discussion

In the present study, two variants of a behavioral treatment program significantly improved the quality of marriages for couples

who received it. The findings replicated a prior study (Jacobson, 1977a). In addition, a new behavioral condition, using contracts of the quid pro quo variety, was found to be significantly improved relative to the WL group on all criterion measures. The changes achieved by QPQ couples were comparable to those achieved by GF couples.

A nonspecific control group was included in the present study to shed light on the importance of those factors that were theoretically assumed to be the active ingredients of the treatment program. Despite the evidence provided by the manipulation checks suggesting that the NS group was perceived as credible, on three of the four dependent measures the behavioral groups improved significantly more than did the NS group. Only on one of the self-report measures, the MHS, did differences between behavioral and NS conditions fail to reach statistical significance. The MHS asks couples to rate their degree of happiness on a 5-point scale in regard to 12 categories of married life. The purposes of this questionnaire are obvious to clients; the way to score "happy" is clear. Intuitively it would appear that of the four measures, the MHS is most susceptible to demand characteristics. This might explain why the discrepancies between the NS and the behavioral groups were smaller on this measure than on any other. This hypothesis is supported by the finding that the nonspecific group improved *only* on this measure.

The findings of the present study indicate that improvement on the part of behavioral couples cannot be entirely a function of nonspecific factors. On the basis of the manipulation checks, it seems clear that NS couples' expectancies were equivalent to those of behavioral couples. They were as "confident" and as "optimistic" regarding their treatment and their therapist as were behavioral couples. NS couples perceived their therapists as interested and involved to the same degree as did behavioral couples. This was true despite the differential effectiveness of the procedures; although NS couples were not improving, they seemed to be satisfied that they were receiving a competent treatment program administered by involved, concerned, competent

therapists. These results were corroborated by undergraduate students' ratings of therapy rationales for behavioral and NS conditions.

Among couples receiving a behavioral form of treatment, it seemed to make little difference whether they used GF contracts or QPQ contracts. Although these results cannot be taken as conclusive, they tentatively suggest that the two contracting forms are interchangeable. However, it is still possible that GF contracts are preferable for severely distressed couples, a hypothesis consistent with Weiss et al. (1974). A better test of this hypothesis would be to cross severity of marital distress with type of contract used in a 2×2 design. Weiss et al. might predict a significant interaction such that the contract forms would be equally effective for moderately disturbed couples like those in the present study, but for severely disturbed couples GF contracting would yield more effective results.

One problem with the present study was the necessity of using the principal investigator as one of the therapists. However, the data do not support the hypothesis of therapist bias influencing the results. First, the couples' posttherapy ratings did not suggest that any of the therapists behaved differently depending on which treatment program they were implementing. In addition, the absence of Therapist \times Treatment interactions argues against therapist bias. The performance of the principal investigator as therapist, relative to those of the other two therapists, was if anything slightly greater in the NS condition than in the two behavioral conditions.

Another limitation of the present study was the small number of couples per treatment cell in the analysis of therapist effects. Although in all but one instance, at least two couples were treated by each therapist in each treatment condition (Therapist 2 treated only one couple in the NS condition, and this couple dropped out prior to completing the program), the statistical tests for therapist differences were not powerful. It was also unfortunate that dropouts had to be included in the analysis for therapist effects (see Data Analysis section). To not include such couples would have seriously biased the appearance of therapist performance. The assumption of no

change from pretest to posttest for the two dropouts was felt to be a cautious estimate, since all other couples receiving treatment, even those who were considered treatment failures, changed in the positive direction on at least some of the measures.

Future studies should investigate the effectiveness of behavioral marital therapy with severely distressed couples; especially needed are studies involving couples in which one or both spouses present severe behavioral problems apart from relationship distress. Component analyses should also be undertaken to evaluate the particular contributions of problem solving and contingency contracting to treatment efficacy. There is already some convergent evidence that the communication training component (problem solving) is both necessary and sufficient for positive change (Jacobson, in press-b), as well as growing controversy regarding the importance of contingency contracting in a behavioral approach to marital problems (e.g., Jacobson, in press-a). One admittedly speculative but plausible interpretation of the apparent equivalence of good faith and quid pro quo contracts is that contracting per se is unnecessary. Only systematic research can provide a definitive answer to this question.

Reference Note

1. Margolin, G. A comparison of marital interventions: Behavioral, behavioral-cognitive, and non-directive. Paper presented at the meeting of the Western Psychological Association, Los Angeles, April 1976.

References

- Harrell, J., & Guerney, B. G. Training married couples in conflict negotiation skills. In D. H. L. Olson (Ed.), *Treating relationships*. Lake Mills, Iowa: Graphic, 1976.
- Hops, H., Wills, T. A., Patterson, G. R., & Weiss, R. L. *Marital interaction coding system*. Eugene: University of Oregon and Oregon Research Institute, 1971. (Order from ASIS/NAPS, c/o microfiche Publications, 305 E. 46th Street, New York, New York 10017.)
- Jacobson, N. S. Problem solving and contingency contracting in the treatment of marital discord. *Journal of Consulting and Clinical Psychology* 1977, 45, 92-100. (a)
- Jacobson, N. S. Training couples to solve their mari-

- tal problems: A behavioral approach to relationship discord: I. Problem-solving skills. *International Journal of Family Counseling*, 1977, 4, 22-31. (b)
- Jacobson, N. S. Training couples to solve their marital problems: A behavioral approach to relationship discord: II. Intervention strategies. *International Journal of Family Counseling*, 1977, 5, 20-29. (c)
- Jacobson, N. S. Contingency contracting: Redundancy and caution. *Behavior Therapy*, in press. (a)
- Jacobson, N. S. A review of the research on the effectiveness of marital therapy. In T. Paolino & B. McCrady (Eds.), *Marriage and the treatment of marital disorders from three perspectives: Psychoanalytic, behavioral, and systems theory*. New York: Bruner/Mazel, in press. (b)
- Jacobson, N. S., & Baucom, D. H. Design and assessment of nonspecific control groups in behavior modification research. *Behavior Therapy*, 1977, 8, 708-719.
- Jacobson, N. S., & Martin, B. Behavioral marriage therapy: Current status. *Psychological Bulletin*, 1976, 83, 540-566.
- Lazarus, A. A. *Behavior therapy and beyond*. New York: McGraw-Hill, 1971.
- Locke, H. J., & Wallace, K. M. Short-term marital adjustment and prediction tests: Their reliability and validity. *Journal of Marriage and Family Living*, 1959, 21, 251-255.
- Olson, D. H., & Ryder, R. G. Inventory of marital conflicts (IMC): An experimental interaction procedure. *Journal of Marriage and the Family*, 1970, 32, 443-448.
- Stuart, R. B., & Stuart, F. *Marital Pre-counseling Inventory*. Champaign, Ill.: Research Press, 1973.
- Tsoi-Hoshmand, L. Marital therapy: An integrative behavioral-learning model. *Journal of Marriage and Family Counseling*, 1976, 2, 179-192.
- Weiss, R. L., Birchler, G. R., & Vincent, J. P. Contractual models for negotiation training in marital dyads. *Journal of Marriage and the Family*, 1974, 36, 321-331.
- Weiss, R. L., Hops, H., & Patterson, G. R. A framework for conceptualizing marital conflict, a technology for altering it, some data for evaluating it. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology concepts and practice*. Champaign, Ill.: Research Press, 1973.

Received February 22, 1977 ■

Toward the Assessment of Social Competence

Robert W. Levenson
Indiana University

John M. Gottman
University of Illinois

Two studies directed toward development and validation of a self-report measure of social competence in dating and assertion situations are described. An 18-item questionnaire consisting of items that assessed the likelihood of certain specific behaviors occurring and the degree of discomfort and expected incompetence in specific situations was derived. This questionnaire discriminated between client and normal populations and between clients with dating and assertion problems, has psychometric properties of reliability and validity, and measures differential improvement following a variety of 8-week intervention programs.

There has been a great deal of recent interest in social skills training, which has been extended from the skill of refusing unreasonable requests (McFall & Lillesand, 1971; McFall & Marston, 1970; McFall & Twentyman, 1973) to more general assertion skills by a number of investigators (Eisler, Hersen, & Miller, 1973; Hersen, Eisler, & Miller, 1973).

The social skills training literature has also expanded to include general social skills training for lower-income clients in mental health centers (Goldstein, 1973), male psychiatric inpatients (Goldsmith & McFall, 1975), and dating skills (Curran, 1975; Curran & Gilbert, 1975; Glass, Gottman, & Shmurak, 1976; Twentyman & McFall, 1975).

In a recent review of social skills training as applied to heterosexual social anxiety, Curran (1977) reviewed 13 studies, concluding that a major issue in the social skills training literature is the assessment of social skills. He noted that "little data exist with regard to the psychometric properties and construct validity of most of the instruments used in previous heterosexual-social anxiety research" (p. 154). Goldfried and Linehan (1977) called for mea-

sures that demonstrate content validity by empirical generation of a content domain (rather than relying on face validity) with attention to the situational context of the behavioral referents assessed. They suggested that discriminant validity studies that demonstrate the separateness of two behavioral concepts will clarify the conceptual ambiguity in behavioral concepts such as assertion.

The present series of investigations is an attempt to develop a self-report assessment measure of social competence that has demonstrated psychometric properties of reliability and validity. Despite the fact that there is a general suspicion of all self-report measures among behavioral scientists, recent research has indicated that under certain specific conditions self-report measures may meet psychometric standards of reliability and validity (Goldfried & Kent, 1972).

Mischel's (1968) review of personality assessment literature led him to conclude that although observation of past behavior in situations with similar role requirements is the best predictor of future behavior in a specific situation, the next best predictor of future behavior is obtained from self-predictions. Furthermore, the research investigations of McFall and his associates have found that although global self-assessments of competence do not relate well to judges' ratings of tapes of behavioral role-playing assessment, self-reports of discomfort and incompetence in *specific* situations (as measured by the Conflict

The authors wish to thank Dave Schlundt, John Embry, Nancy Levin, and Jim Barrett for their help in carrying out this research.

Requests for reprints should be sent to John M. Gottman, Department of Psychology, University of Illinois, 505 East Green Street, Champaign, Illinois 61820.

Resolution Inventory) do correlate well with behavioral assessments. For example, McFall and Lillesand (1971) wrote:

The results obtained on the assertive score, nonassertive score, and difference score measures [of the Conflict Resolution Inventory], all of which assessed responses in specific refusal situations, were in sharp contrast to the nonspecific effects obtained on the global measure. (pp. 316-317).

This general finding has been replicated by other investigators (e.g., Schwartz & Gottman, 1976). In a social skills training study with male psychiatric inpatients, Clark (1975) used a global self-assessment of improvement, a situationally specific self-assessment, and a behavioral role-playing assessment. The control group, which received didactic lectures, showed no improvement on the behavioral assessment measure and no improvement on the situationally specific assessment measure but did show improvement on the global self-assessment measure. The social skills training group showed improvement on all three measures. There is thus some initial evidence suggesting that a situationally specific self-report measure of social competence would have validity with respect to laboratory role-playing assessments.

The current investigation requires a self-report measure of social competence to demonstrate several specific kinds of validity. First, it must discriminate between competent and incompetent populations, with competence independently defined. Second, it must discriminate among specific types of social incompetence; for example, nonassertive subjects should show a different scale pattern profile than subjects with heterosexual dating problems. Third, in cases in which treatment is used, the self-report measure must predict differential improvement in treatments designed for the amelioration of specific problems. For example, nonassertive subjects should generally improve on assertion items but not on dating items, compared to dating-problem subjects, who should improve on dating but not assertion items, compared to nonassertive subjects. This latter criterion of validity is dependent on intervention programs that target specific skills for training, and will probably not be met to the extent that dating skills training programs and assertion training programs overlap in the skills they teach.

The present series of investigations was undertaken to design a self-report measure that meets the three criteria of validity described above, as well as internal consistency and test-retest reliabilities. The present investigations also followed the recommendation of Goldfried and D'Zurilla (1969) in empirically constructing a domain of problematic social situations. From this domain, items that involved two specific self-reports were constructed: (a) self-report of discomfort or incompetence—dimensions that have shown validity with behavioral assessments in McFall's Conflict Resolution Inventory—and (b) self-report of the likelihood of engaging in specific behaviors. Items were selected from the larger domain in the two subdomains of assertion and heterosexual dating. A series of reliability and validity studies were undertaken using these items.

Study 1

Method

Subjects

During the second week of the fall 1976 semester, a notice announcing the availability of social skills training programs for students having problems in dating and assertion situations was placed in the student newspaper and posted on dormitory bulletin boards. The approximately 200 students who responded to the notice were mailed a package that included information about the training programs and three questionnaires (described below). Respondents were requested to complete the three questionnaires and return them along with a \$5 deposit if they wished to be included in a training program. They were informed that the deposit would be refunded when they completed a second set of questionnaires at the end of the program. When registration was terminated 3 weeks after the notice first appeared, 92 students had completed the pretest materials, and these students became the "client" population for the study.

At the same time, a group of 69 students who had not signed up for the training program were recruited from the introductory psychology classes and were given the complete set of questionnaires. These students were the "normal" population for the first experiment.

Procedure

Three questionnaires were administered to the client and normal populations: (a) a situations questionnaire, (b) a behavior inventory, and (c) a symptom checklist. A description of these questionnaires follows.

Situations questionnaire (40 items). A domain of items was generated by eight undergraduates (four

Table 1
Client and Normal Subject's Pretest Data on Overall and Subscale Scores

Test	<i>M</i>		<i>F</i> (1, 157)
	Client	Normal	
Situations			
Overall	3.0	3.6	55.575***
Refusal	3.5	3.7	4.127*
Getting What You Want	3.1	3.4	5.521**
Expressing Feeling	2.8	3.6	56.364***
Requesting Behavior Change	3.2	3.7	17.494***
Formal Situations	2.8	3.5	53.734***
Conversation Skills	2.6	3.6	75.947***
Close Interpersonal Situations	3.3	3.7	25.177***
Dating	2.6	3.3	40.133***
Behavior			
Overall	1.0	2.5	84.222***
Friendship	1.9	2.7	71.480***
Self-confidence	1.9	2.3	32.613***
Assertiveness	2.4	2.6	9.171***
Intimacy	1.9	2.3	32.290***
Dating	1.8	2.5	54.271***
Symptom			
Overall	2.2	1.8	20.917***

* $p < .041$.

** $p < .019$.

*** $p < .001$.

males and four females) who signed up for a topical seminar on interviewing. Each member of the seminar interviewed 10 undergraduates and obtained a description of four social situations that the interviewee had recently found to be "difficult to handle." A description of each situation, written by the interviewer, summarized the situational context, the roles of the principal characters in the situation, the action, and the time of key difficulty that preceded a response demanded of the interviewee. The original list of 320 situations was used to generate 97 nonredundant items that could be potentially relevant to both sexes and that struck a balance between being overly general or overly specific. The items were sorted into seven a priori scales by the content of the task posed by the situation: (a) refusing unreasonable requests, (b) getting what you want, (c) expressing how you feel, (d) requesting behavior change from someone, (e) dealing with formal situations (such as a dinner party), (f) initiating and continuing conversations, and (g) dating situations (such as asking for a date and getting close to someone of the opposite sex).

Durham (Note 1) tested these a priori scales with 126 undergraduates. He used three phrasings of the self-report question: (a) a phrasing that confounded discomfort with incompetence,¹ (b) a discomfort phrasing, and (c) an incompetence phrasing. The confounded phrasing showed the best a priori scale test-retest reliabilities (.75) between administrations 3

weeks apart and the best Cronbach alpha coefficient (.97) and split-half reliability coefficient (.94). Using an item analysis of the correlation of items with a priori subscale totals, Durham reduced the original 97-item questionnaire to 40 items. Durham also conducted analyses of selected subject characteristics and found no differences between subjects' scores as a function of sex, year in college, or marital status.

The following excerpt from the social situations questionnaire illustrates the format used:

After each situation, circle one of the numbers from 1 to 5 which best describes you using the following scale:

- 1 = I would be so uncomfortable and so unable to handle this situation that I would avoid it if possible.
- 2 = I would feel very uncomfortable and would have a lot of difficulty handling this situation.
- 3 = I would feel somewhat uncomfortable and would have some difficulty in handling this situation.
- 4 = I would feel quite comfortable and would be able to handle this situation fairly well.

¹ This was the phrasing used in the Conflict Resolution Inventory.

5 = I would feel very comfortable and be able to handle this situation very well.

Your friend's relatives invite you over for dinner. You accept, then begin to feel nervous about making a good impression. You arrive at their house, and everyone sits down to talk before dinner. One of the relatives smiles at you and seems to expect you to say something. 1 2 3 4 5

Behavior inventory (26 items). Construction of the behavior inventory was considerably less formal than that of the social situations questionnaire. Five a priori subscales were established based on five social skills training groups that had been offered by clinical psychology graduate students supervised by us during the spring 1975 semester. The five groups were (a) friendship, (b) self-confidence, (c) assertion, (d) intimacy, and (e) dating. Behaviors that were seen as being particularly difficult for participants in each group were converted into items on the inventory. The inventory was constructed to assess the likelihood of a respondent to exhibit these behaviors. The following excerpt illustrates the nature of the inventory, with examples from the self-confidence, assertion, and dating subscales.

How much were you bothered by:

	Not at all	A little bit	Moderately	Quite a bit	Extremely
Headaches	1	2	3	4	5
Nervousness or shakiness inside	1	2	3	4	5
Trouble remembering things	1	2	3	4	5

Results and Discussion

The presentation of the results is divided into three sections: One section is related to the first validity claim, namely, discrimination of clients from nonclients; one section is related to the second validity claim, namely discrimination of assertion clients from dating skills clients; and a third section is related to psychometric properties of the measures.

Clients and Nonclients

Data were analyzed separately for a priori subscales, individual items (on the situations questionnaire and the behavior inventory) for overall average item score for each questionnaire and for the total symptom checklist score. These data were analyzed in a two-way (clients vs. normals) analysis of variance. An unweighted means solution was used because of the unequal sample sizes. These analyses revealed that clients had greater difficulty on all subscale scores for both the situations

We are interested in finding out something about the likelihood of your acting in certain ways. Below you will find a list of specific behaviors you may or may not exhibit. Use the following rating scale:

1	2	3	4
I never do this	I sometimes do this	I often do this	I do this almost always

Now after each of the items on the following list, place the number which best indicates the likelihood of your behaving in that way. Be as objective as possible.

Volunteer to do something where there is a good chance you might fail. _____

Say "no" when you feel like it. _____

Start a conversation with a member of the opposite sex you would like to date. _____

Symptom checklist (90 items). A questionnaire normally used with hospital inpatients² was adopted for use. Items on this questionnaire reflect anxiety, depression, and somatic symptoms. Subjects rated each item to the extent they are troubled by that problem on a 1 to 5 scale. The following excerpt illustrates the nature of the checklist:

questionnaire and the behavior inventory, greater difficulty on the overall average item score on all three questionnaires, and greater difficulty on 21 of 26 items on the behavior inventory and 35 of 40 items on the situations questionnaire. In all cases, the significance level of these differences was less than .05. Table 1 presents means, *F* ratios, and *p* levels for the subscale and overall average item scores for the clients and normals. The *F* ratios indicate a considerable degree of discriminative power. Note that the means refer to the item scale values described above for each questionnaire and that *smaller* numbers indicate greater difficulty on the situations questionnaire and behavior inventories, whereas *larger* numbers indicate greater difficulty on the symptom checklist.

² This symptom checklist was used as part of a standard clinical intake procedure by the Illinois State Psychiatric Institute and the Family Institute of Chicago.

Table 2
Dating and Assertion Subjects' Pretest Data on Overall and Subscale Scores

Test	<i>M</i>		<i>F</i> (1, 86)	<i>p</i> <
	Dating	Assertion		
Situations				
Overall	3.0	3.0	—	
Refusal	3.6	3.5	2.674	.102
Getting What You Want	3.2	3.0	2.818	.093
Expressing Feelings	2.9	2.9	—	
Requesting Behavior Change	3.3	3.2	—	
Formal Situations	2.9	2.8	1.214	.273
Conversational Skills	2.5	2.8	1.777	.183
Close Interpersonal Situations	3.2	3.4	—	
Dating	2.4	2.8	8.224	.005
Behavior				
Overall	2.0	2.0	—	
Friendship	1.9	2.0	—	
Self-confidence	2.0	1.8	8.635	.005
Assertiveness	2.5	2.3	4.684	.030
Intimacy	1.8	2.0	—	
Dating	1.6	2.0	11.934	.001
Symptom				
Overall	2.1	2.3	3.420	.064

The results clearly indicate that students who signed up for social skills training reported much greater difficulty across the range of social dimensions measured by our instruments than did normal students. Interestingly, they also reported a greater prevalence of "psychiatric" and somatic symptoms. A picture emerged of a subpopulation that may present itself as generally less socially competent and more problem ridden than its peers.

The results offer some initial validation of the a priori subscales used in the situations and behavior questionnaires: The subscales successfully discriminated between client and normal populations. The first validity criterion was therefore satisfied.

Assertion Clients and Dating Skills Clients

This analysis was carried out using the same data, except that only the 92 clients were used: the assertion groups clients ($n = 46$) and the dating skills clients ($n = 46$).

Two-way (dating subjects vs. assertion subjects) analyses of variance were performed for overall questionnaire scores, a priori subscales, and individual items. Results indicate

that dating subjects showed significantly greater difficulty as compared to assertion subjects on the dating subscales of both the situations questionnaire and the behavior inventory. Assertion subjects showed significantly greater difficulty on the "self-confidence" and "assertiveness" subscales of the behavior inventory. Means, *F* ratios, and *p* levels are presented for these differences in Table 2. Analysis of the individual items revealed that 8 of 40 items on the situations questionnaire and 8 of 26 items on the behavior inventory significantly differentiated dating subjects and assertion subjects at $p < .05$.

Knowing the training program for which clients had registered allowed a second empirical test of the validity of several of our subscales. The results of this experiment indicate that clients with dating and assertion problems tend to score accordingly on dating-related and assertion-related subscales. Moreover, the use of two different kinds of self-report measures (i.e., the situations questionnaire and behavior inventory), and the tendency of clients to score appropriately on both, provided us with convergent evidence that true differences existed between the dating and

assertion subpopulations and that these were measurable independently using self-report measures.

To enhance the validity of the dating and assertion scales, we decided to focus on the dating and assertion subpopulations, to concentrate on developing one questionnaire containing only the dating and assertion subscales, and to subject this new questionnaire to standard reliability tests prior to continuing with additional validation procedures.

Psychometric Properties

An 18-item questionnaire was developed with a 9-item dating subscale and a 9-item assertion subscale. This new questionnaire was tested scalewise for internal consistency and test-retest reliability. In addition, previous validity tests for discriminating clients versus normals and dating versus assertion problems were recomputed using these new subscales.

The original 92 clients (46 dating, 46 assertion) and 69 normals were studied again for computing internal consistency and for performing concurrent validity checks. Seventy additional subjects who had not registered for the training programs were recruited from the introductory psychology classes to serve as a sample for performing a test-retest reliability analysis.

The original 26-item behavior inventory and 40-item situations questionnaire were transformed into an 18-item questionnaire by selecting only those items that both successfully discriminated clients from normals and successfully discriminated dating clients from assertion clients. Of the 18 items that met these criteria, dating clients indicated having greater difficulty with 9 of the items (5 from the original behavior inventory and 4 from the original situations questionnaire), whereas assertion clients indicated greater difficulty with the other 9 (4 from the original behavior inventory and 5 from the original situations questionnaire). Thus, these sets of items became our 9-item dating and assertion subscales (see Appendix), which were tested for their psychometric properties as follows: (a) A Cronbach alpha was computed for assessing the internal consistency of the dating and assertion subscales using the data from the

original clients and normals; (b) comparisons of clients versus normals and of dating clients versus assertion clients were made on the two subscales using the original client and normal sample; (c) to assess test-retest reliability 6 weeks prior to the end of the semester 44 normal subjects were administered the original test battery. An additional 30 normal subjects took the test battery 4 weeks later. All 70 subjects took the battery again 2 weeks later. Usable data were obtained from 28 subjects for the 2-week test-retest interval and from an independent group of 39 subjects for the 6-week interval.

Analysis of internal consistency yielded a Cronbach alpha of .92 for the dating subscale and an alpha of .85 for the assertion scale. Concurrent discriminant validity analyses revealed clients to have significantly greater difficulty than normals on both the dating subscale, $F(1, 159) = 52.60$, $p < .001$, and the assertion subscale, $F(1, 159) = 34.33$, $p < .001$. Dating clients had more difficulty than assertion clients on the dating subscale, $F(1, 86) = 17.55$, $p < .001$. Assertion clients had more difficulty than dating clients on the assertion subscale, $F(1, 86) = 21.00$, $p < .001$.

To assess test-retest change, a 2×2 (2 Week vs. 6 Week \times Pretest vs. Posttest) analysis of variance was computed for the 2-week and 6-week groups. The results indicated no change at retesting at either interval for either the dating subscale or the assertion subscale. The test-retest correlations for both subscales at both testing intervals ($ns = 28$ and 39, respectively) were: For dating at 2 and 6 weeks, $rs = .71$ and $.62$. For assertion, $rs = .71$ and $.70$ ($p < .001$).

The results indicate that the dating and assertion subscales have demonstrable psychometric qualities of reliability and validity. Of particular interest was the finding that the scales had internal consistency despite the fact that items were selected on the basis of their ability to discriminate between populations. This suggests that the scale items are in fact measuring the same dimension and that this dimension is one for which salient differences do exist between the populations in question.

The test-retest experiment was performed to determine whether the subscales would fluctuate greatly over the measurement periods

in question. Especially important for Study 2 was the determination of whether significant changes in self-report of dating and assertion difficulties would occur at the end of semester approached. On the basis of our findings, there is no reason to expect these kinds of difficulties to spontaneously increase or decrease over the course of our testing intervals. However, these test-retest data were obtained using normal subjects, and their applicability to client populations was not tested.

Study 2

In this section the results of an 8-week intervention directed toward amelioration of specific social skills problems is presented. This intervention was used to test the abilities of our instrument to measure differential changes as a function of the type of social skills training program.

Method

Procedure

The 46 dating clients who had completed the pretest materials in Experiment 1 were assigned to one of three treatment conditions: (a) group meeting ($n = 11$), (b) self-help manual plus consultant ($n = 11$), or (c) self-help manual ($n = 24$).

In a similar manner, the 46 assertion clients were assigned to either group meeting ($n = 8$), self-help manual plus consultant ($n = 12$), or self-help manual ($n = 26$) conditions. A description of the three treatment conditions follows.

Group meeting. Clients assigned to this condition attended weekly 90-minute sessions for 8 weeks under the leadership of male and female cotherapists. The focus of these groups was on behavioral rehearsal, role-playing, and skill acquisition exercises.

Self-help manual plus consultant. Original manuals were written that contained information and exercises relevant to assertion skills and dating skills.³ These manuals were divided into eight sections, with each containing information and exercises for 1 week.

In addition to the manual, clients in this condition were assigned an undergraduate "consultant" who met with the client at the start of the 8-week period, called them periodically to check on their progress, and was available for phone consultation if the client so desired.

Self-help manual. Clients in this condition received the appropriate self-help manual as in the previous condition but were not assigned a consultant.

At the end of the 8-week period, all clients in all conditions were mailed a package of posttest questionnaires and were reminded that their \$5 deposits would be refunded as soon as the materials were completed and returned.

Table 3

Pretest and Posttest Subscale Scores for Dating and Assertion Clients

Group	Dating	Assertion
Dating		
Pre (46)	1.95	2.91
Post (38)	2.41	3.02
t	5.55*	1.34
Assertion		
Pre (46)	2.45	2.48
Post (35)	2.76	3.02
t	3.75*	6.51*

Note. Numbers in parentheses are n s.

* $p < .001$.

Results and Discussion

We were able to obtain a fairly high rate of return from clients completing the program for our posttest materials. There was also a small number of clients who chose to drop out of the program prior to its completion. The overall return rate was 79%, and by treatment was group meeting (89%), self-help manual plus consultant (78%), and self-help manual (76%).

Data obtained from these clients were analyzed in a 2×2 (Dating Clients versus Assertion Clients \times Pretest versus Posttest) analysis of variance for the dating and assertion subscales. Results indicated a significant pretest versus posttest main effect for both the dating subscale, $F(1, 67) = 48.31$, $p < .001$, and the assertion subscale, $F(1, 67) = 37.87$, $p < .001$. Significant Client \times Test interactions were obtained for the dating subscale, $F(1, 67) = 4.40$, $p = .037$, and for the assertion subscale, $F(1, 67) = 21.11$, $p < .001$.

Of greatest interest for the present investigation were the data concerning changes on the dating and assertion subscales for clients working in dating and assertion training programs. Pretest and posttest means for these clients and subscales are presented in Table 3. These results indicate that significant improvement occurred only for the dating subscales for dating clients. Assertion clients improved

³ The authors are extremely grateful to John Embry, Jennifer Parkhurst, and David Schlundt who helped write and edit the manuals.

on both subscales but showed more improvement on the assertion scale than on the dating scale.

The results obtained from Study 2 indicate measurable improvement in both client populations over the 8-week period. Despite the fact that a no-treatment control was not included, it can be argued for several reasons that these changes are most readily attributable to the interventions that occurred during this period. First, the most pronounced change occurred on the subscale related to the targeted problem. This was especially true for the dating clients, who showed no change on the assertion subscale. In addition, test-retest data on normal subjects in Study 1 over a similar time period indicated no change on either subscale. There is little reason to expect that client populations would spontaneously improve over this time period.

Study 2 provides an extension of the utility of the dating and assertion subscales. Prior to initiating these pretreatment versus post-treatment comparisons, we were not optimistic about the likelihood that these subscales would be useful for measuring change following an 8-week intervention. It had seemed to be the case that the utility of a scale to register changes on a personality dimension was quite independent of its ability to satisfy static criteria of reliability and validity. The additional expectation of differential changes as a function of specific types of intervention thus serves as an additional validity check. The discriminant validity of the dating and assertion scales may clarify the frequent ambiguity inherent in the behavioral concept of social skills; it would seem reasonable to hypothesize that social competence consists of a set of relatively independent skills.

This article is a step toward the assessment of specific aspects of social competence. We should add that these two scales should be used cautiously; the two questionnaires contain a narrow sampling of items from a larger domain (cf. Durham, Note 1) and should not be equated with social competence. We also stress the limitation of this article in only using self-report measures in the validation procedure. Still to be demonstrated is that these measures correlate with relevant extralaboratory criteria and with measures obtained by coding behavior samples.

Our primary interest was the construction of measures that successfully differentiate people who have a given difficulty from those who do not, that discriminate among people who have different kinds of related difficulties, and that indicate change in the level of this difficulty differentially as a function of the treatment received. The results of the studies presented here indicate that one kind of self-report measure that satisfies all of these criteria can be constructed by assessing the likelihood of certain behaviors occurring and the degree of discomfort and expected incompetence in specific situations. A useful assessment device was thus constructed from items that combined behavioral specificity with the phenomenology of expected difficulty and discomfort.

Reference Note

1. Durham, R. *The social questionnaire: A new measure of social competence among college students*. Unpublished manuscript, Indiana University, 1976.

References

- Clark, K. W. Evaluation of a group social skills training program with psychiatric inpatients: Training Viet Nam era veterans in assertion, heterosexual and job interview skills. (Doctoral dissertation, University of Wisconsin, 1974). *Dissertation Abstracts International*, 1975, 35, 4642B. (University Microfilms No. 74-28, 795).
- Curran, J. P. An evaluation of a skills training program and a systematic desensitization program in reducing dating anxiety. *Behaviour Research and Therapy*, 1975, 13, 65-68.
- Curran, J. P. Skills training as an approach to the treatment of heterosexual-social anxiety: A review. *Psychological Bulletin*, 1977, 84, 140-157.
- Curran, J. P., & Gilbert, F. S. A test of the relative effectiveness of a systematic desensitization program and an interpersonal skills training program with date-anxious subjects. *Behavior Therapy*, 1975, 6, 510-521.
- Eisler, R. M., Hersen, M., & Miller, P. M. Effects of modeling on components of assertive behavior. *Journal of Behavior Therapy and Experimental Psychiatry*, 1973, 4, 1-6.
- Glass, C. R., Gottman, J. M., & Shmurak, S. S. Response acquisition and cognitive self-statement modification approaches to dating skills training. *Journal of Counseling Psychology*, 1976, 23, 520-526.
- Goldfried, M. R., & D'Zurilla, T. J. A behavioral-analytic model for assessing competence. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 1). New York: Academic Press, 1969.

- Goldfried, M. R., & Kent, R. N. Traditional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin*, 1972, 77, 409-420.
- Goldfried, M. R., & Linehan, M. M. Basic issues in behavioral assessment. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Goldsmith, J. B., & McFall, R. M. Development and evaluation of an interpersonal skill-training program for psychiatric inpatients. *Journal of Abnormal Psychology*, 1975, 84, 51-58.
- Goldstein, A. P. *Structured learning therapy: Toward a psychotherapy for the poor*. New York: Academic Press, 1973.
- Hersen, M., Eisler, R. M., & Miller, P. M. Effects of practice, instructions, and modeling on components of assertive behavior. *Behaviour Research and Therapy*, 1973, 11, 443-451.

- McFall, R. M., & Lillesand, D. B. Behavior rehearsal with modeling and coaching in assertion training. *Journal of Abnormal Psychology*, 1971, 77, 313-323.
- McFall, R. M., & Marston, A. R. An experimental investigation of behavior rehearsal in assertive training. *Journal of Abnormal Psychology*, 1970, 76, 295-303.
- McFall, R. M., & Twentyman, T. T. Four experiments on the relative contributions of rehearsal, modeling, and coaching to assertion training. *Journal of Abnormal Psychology*, 1973, 81, 199-218.
- Mischel, W. *Personality and assessment*. New York: Wiley, 1968.
- Schwartz, R. M., & Gottman, J. M. Toward a task analysis of assertive behavior. *Journal of Consulting and Clinical Psychology*, 1976, 44, 910-920.
- Twentyman, C. T., & McFall, R. M. Behavioral training of social skills in shy males. *Journal of Consulting and Clinical Psychology*, 1975, 43, 384-395.

Appendix

Dating and Assertion Questionnaire (18 items)

We are interested in finding out something about the likelihood of your acting in certain ways. Below you will find a list of specific behaviors you may or may not exhibit. Use the following rating scale:

- 1 = I never do this
- 2 = I sometimes do this
- 3 = I often do this
- 4 = I do this almost always

Now after each of the items on the following list, place the number which best indicates the likelihood of your behaving in that way. Be as objective as possible. [Subscale loadings for items are indicated in parentheses. A = assertion subscale; D = dating subscale.]

1. Stand up for your rights (A)
2. Maintain a long conversation with a member of the opposite sex (D)
3. Be confident in your ability to succeed in a situation in which you have to demonstrate your competence (A)
4. Say "no" when you feel like it (A)
5. Get a second date with someone you have dated once (D)
6. Assume a role of leadership (A)
7. Be able to accurately sense how a member of the opposite sex feels about you (D)
8. Have an intimate emotional relationship with a member of the opposite sex (D)

9. Have an intimate physical relationship with a member of the opposite sex (D)

The following questions describe a variety of social situations that you might encounter. In each situation you may feel "put on the spot." Some situations may be familiar to you, and others may not. We'd like you to read each situation and try to imagine yourself actually in the situation. The more vividly you get a mental picture and place yourself into the situation, the better.

After each situation circle one of the numbers from 1 to 5 which best describes you using the following scale:

- 1 = I would be so uncomfortable and so unable to handle this situation that I would avoid it if possible.
- 2 = I would feel very uncomfortable and would have a lot of difficulty handling this situation.
- 3 = I would feel somewhat uncomfortable and would have some difficulty in handling this situation.
- 4 = I would feel quite comfortable and would be able to handle this situation fairly well.
- 5 = I would feel very comfortable and be able to handle this situation very well.

1. You're waiting patiently in line at the checkout when a couple of people cut right in front of you. You feel really annoyed and want to tell them to wait their turn at the back of

the line. One of them says, "Look, you don't mind do you? But we're in a terrible hurry."

1 2 3 4 5 (A)

2. You have enjoyed this date and would like to see your date again. The evening is coming to a close and you decide to say something.

1 2 3 4 5 (D)

3. You are talking to a professor about dropping a class. You explain your situation, which you fabricate slightly for effect. Looking at his grade book the professor comments that you are pretty far behind. You go into greater detail about why you are behind and why you'd like to be allowed to withdraw from his class. He then says, "I'm sorry, but it's against university policy to let you withdraw this late in the semester."

1 2 3 4 5 (A)

4. You meet someone you don't know very well but are attracted to. You want to ask them out for a date.

1 2 3 4 5 (D)

5. You meet someone of the opposite sex at lunch and have a very enjoyable conversation. You'd like to get together again and decide to say something.

1 2 3 4 5 (D)

6. Your roommate has several obnoxious traits that upset you very much. So far, you

have mentioned them once or twice, but no noticeable changes have occurred. You still have 3 months left to live together. You decide to say something.

1 2 3 4 5 (A)

7. You're with a small group of people who you don't know too well. Most of them are expressing a point of view that you disagree with. You'd like to state your opinion even if it means you'll probably be in the minority.

1 2 3 4 5 (A)

8. You go to a party where you don't know many people. Someone of the opposite sex approaches you and introduces herself. You want to start a conversation and get to know him/her.

1 2 3 4 5 (D)

9. You are trying to make an appointment with the dean. You are talking to his secretary face-to-face. She asks you what division you are in and when you tell her, she starts asking you questions about the nature of your problem. You inquire as to why she is asking all these questions and she replies very snobbishly that she is the person who decides if your problem is important enough to warrant an audience with the dean. You decide to say something.

1 2 3 4 5 (A)

Received February 28, 1977 ■

Behavior Modification for Obesity: The Evaluation of Exercise, Contingency Management, and Program Adherence

Peter M. Stalonas, Jr.
University of Rochester

William G. Johnson
Department of Psychiatry and Human Behavior
University of Mississippi Medical Center

Maryann Christ
University of Rochester

This study investigated crucial aspects of behavioral programs for obesity including (a) the assumption that subjects actually engage in requested behaviors and that these behaviors mediate weight loss, (b) the effect of exercise on weight loss, and (c) the problem of long-term maintenance and generalization to the clinically obese. Exercise and self-managed contingency components were compared in a 2×2 factorial design on 44 obese subjects and were evaluated after 10 weeks of treatment and 3-month and 1-year follow-ups. Significant weight loss was observed for all groups at program termination ($p < .001$) and the 3-month follow-up ($p < .001$), with only those exposed to exercise and/or contingency management maintaining weight loss after 1 year. There were no main effects or interactions at program termination or at the 3-month follow-up. However, the influence of exercise at the 1-year follow-up was noticeable ($p < .10$). Assessment of program adherence indicated that subjects engaged in program behaviors, yet only 1 of 10 such behaviors was related to weight loss.

Recent reviews of behavior modification for obesity support the superiority of this approach relative to other treatments (Leon, 1976; Stunkard, 1975). However, several important issues regarding the evaluation of the behavioral approach remain. These include (a) a common and perhaps erroneous assumption that subjects actually engage in program behaviors and that these behaviors mediate weight loss, (b) the failure to evaluate the influence of exercise, (c) long-term evaluation, and (d) the use of clinical as opposed to student populations.

The first issue of subject compliance is highlighted by a survey of over 20 recent reports

on the behavioral management of obesity (Johnson & Stalonas, Note 1). Only 3 of these reports supplied information relevant to whether subjects actually engaged in the behavioral changes suggested. For example, studies by Hall, Hall, Hanson, and Borden (1974); Horan, Baker, Hoffman, and Shute (1975); Jeffrey (1974); and Mahoney, Moura, and Wade (1973), while demonstrating weight loss, either provided no information on program adherence or were limited to post hoc reports (Bellack, Rozenky, & Schwartz, 1974; Mahoney, 1974) or pre-post questionnaires on eating habits (Wollersheim, 1970). Thus, information supporting the claim that participants actually use behavioral strategies is notably lacking in the published literature. Investigators may be incorrectly inferring the operation of independent variables embodied in these programs on the basis of weight change alone (Mahoney, 1975).

A second issue in the treatment of obesity is the influence of exercise. Low levels of energy expenditure are typical of the obese, and Mayer

Senior authorship is shared by the first two authors. Portions of this article were presented at the annual meetings of the Southeastern Psychological Association, New Orleans, March 1976, and the American Psychological Association, Washington, D.C., September 1976. Requests for reprints should be sent to William G. Johnson, Department of Psychiatry and Human Behavior, University of Mississippi Medical Center, Jackson, Mississippi 39216.

(1968), among others, indicated that this relative inactivity is a major factor in the development of obesity. Exercise not only increases caloric expenditure and the metabolism of fat, but it also decreases appetite, aids in cardiovascular conditioning, and generally promotes a sense of psychological well-being (Björntorp, 1976; Horton, 1974). Thus, the direct health benefits of exercise are obvious and strongly implicated as an important adjunct to weight reduction.

In spite of its appeal, the contribution of regular exercise to weight loss has been evaluated in only one study. In this study, Harris and Hallbauer (1973) found no difference at the termination of a 12-week program between *controlled eating* and *controlled eating plus exercise* groups. However, the group that exercised lost significantly more weight at a 7-month follow-up than the comparison group. This study stands in virtual isolation, and a replication of the exercise effect is indicated particularly in light of its potential contribution to the maintenance of weight loss.

The third issue is the evaluation of subjects over an extended period of time. The overwhelming majority of studies on weight reduction follow subjects for 3 months or less. For the clinically obese who are often 50% over their ideal weight, long-term follow-up of at least 1 year is necessary to consider program efficacy.

Finally, past research on obesity has been largely limited to subjects in college settings. Evaluation of the effects of a behavioral weight loss intervention on community members of varying ages and socioeconomic classes is necessary if we are to address the generalized utility of such programs.

The present study confronted these crucial aspects of behavioral approaches to weight reduction by attempting to (a) ascertain the relationship between ongoing reports of program adherence and weight loss; (b) assess the effectiveness of exercise; (c) plan long-term evaluation; and (d) use a clinical, nonstudent population.

Method

Subjects

Forty-eight subjects responded to local advertisements of a free program for weight loss. During the

first 3 weeks of the program, 4 subjects dropped out due to scheduling or health problems, leaving 44 subjects who met the following requirements: (a) at least 15% overweight based on figures tabulated by the Metropolitan Life Insurance Company (1959), (b) free of complicating health disorders and medication, and (c) no prior experience with formal behavioral programs for weight loss.

Subjects averaged 31.5 years of age (range = 16-62), 181.1 pounds (82.1 kg) (range = 130-275 pounds) (59.0-124.7 kg), and 40.2% overweight (range = 15%-112%). There were 37 female and 7 male participants who were from socioeconomic status Levels I to IV (Hollingshead & Redlich, 1958) with 2, 19, 21, and 2 members per level, respectively.

Procedure

Subjects were randomly assigned to four groups (matched on age, sex, absolute weight, and overweight) in a 2×2 factorial design. Also, baseline levels of activity were assessed, and mean expenditures for the four groups were equivalent. All groups received a similar weight loss program (P) with the following exceptions: One group received the addition of both exercise and contingency components (PEC), a second group received the addition of just exercise (PE), a third group received the addition of just a contingency component (PC), and the final group was exposed to the basic program alone (P).

The basic program used in all four groups consisted of 10 written lessons (Johnson & Stalonas, in press) that describe a sequence of behavioral tasks including monitoring all information related to eating; advice on the components of a balanced diet; making salient those activities that might inhibit eating; eating at specific times, places, and situations; manipulating elements in the eating chain; imagining aversive stimuli contingent on inappropriate urges to eat; and graphing the use of program behaviors.

The exercise component consisted of specific attempts to increase weekly levels of physical activity. PEC and PE groups were given a list relating activities to caloric expenditures and were encouraged to increase their physical activity over a 10-week period from 150 to 400 kcalories (630-1,680 J) per day above their basal level.

The contingency manipulation consisted of instruction in the use of self-reinforcement for successfully applying the strategies of the program. Subjects in the PEC and PC groups compiled lists of activities that could serve as self-administered rewards. In the third week, these subjects completed daily checklists in which program activities were converted into points that could be exchanged for daily (e.g., reading a chapter of a book) and weekly rewards (e.g., buying a new dress).

Group Meetings

The four groups of subjects attended 10 weekly 1-hour sessions, which were conducted by two experi-

enced coleaders. To maximize attendance, subjects deposited \$10, of which \$1 was returned each week. This procedure resulted in an average attendance for all subjects of nine meetings.

The first part of each meeting was directed toward an extensive, individual review of performance records, charts, and weights with one of the two therapists. Thereafter, a group discussion ensued regarding progress, areas of difficulty, and suggestions with liberal administrations of verbal encouragement and support. Toward the end of the group meeting, the next lesson was distributed and explained.

Results

There were two main sources of data—weight change and program behaviors. For weight change, data were available for Week 1 through program termination at Week 10, a 3-month follow-up, and a 1-year follow-up. Data regarding program behavior adherence are available for Week 1–Week 10. Of the 44 subjects completing the program, one subject was unavailable at the 3-month follow-up due to pregnancy, and two additional subjects had left the country at the 1-year follow-up.

Weight Loss

Mean weight changes for each of the four groups over three time periods are presented in Table 1. These changes were analyzed in three separate 2×2 analyses of variance for each time period.¹ The main effects of contingency management and exercise and the interactions did not reach significant levels in any analysis. There was a nonsignificant tendency, however, toward a main effect of exercise for Week 1 to the 1-year follow-up, $F(1, 37) = 3.0, p < .10$.

Correlated t tests were performed for each of the four groups separately. As Table 1 reveals, each group lost a significant amount of weight at program termination. ($M = 10.7$ pounds (4.8 kg) for the four groups taken together, $p < .001$.) This weight loss amounts to a decrease of 8% in excess weight.

At the 3-month follow-up, average weight changes for each group remained significant at the .001 level, with the total weight loss averaging 12.50 pounds (5.7 kg). Although there were no significant differences between Week 10 and the 3-month follow-up for any group, most subjects maintained the weight

Table 1
Mean Weight Loss

Group	Week 10	Follow-up	
		3 mo.	1 yr.
PEC	9.5***	12.4***	13.1**
%	10	10	10
PE	13.1***	14.9***	16.3*
%	10	9	9
PC	10.0***	13.4***	11.4***
%	12	12	10
P	10.3***	9.9***	5.2
%	12	12	12

Note. PEC = program with exercise and contingency components; PE = program with just exercise; PC = program with just the contingency component, and P = the basic program alone.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

loss engendered during the program or continued to lose weight.

Data available at the 1-year follow-up revealed a similarly striking yet less consistent pattern of maintained weight loss. As Table 1 illustrates, the three groups exposed to contingency and/or exercise components were able to maintain their weight loss. In contrast, the P group, exposed to only the basic training in stimulus control, displayed an average weight increase of 4.7 pounds (2.1 kg) from the 3-month to the 1-year follow-up.

Program Adherence

The performance of program behaviors was assessed on the basis of weekly reviews of each subject's recordings and graphs. Ratings were made of stimulus control, notebook recordings, self-monitoring with graphs, chaining, timeouts, one portion of food per meal, caloric monitoring, between-meal uncontrolled eating, exercise, and self-administered reinforcement. With the exception of exercise and uncontrolled eating, program behaviors were rated on a 3-point scale based on the number of days per week in which the behavior was performed

¹ Statistical analyses were also performed on percent overweight and Feinstein's (1959) weight reduction index with results virtually identical to those for weight change.

correctly. One point was awarded for 6 or 7 days, $\frac{1}{2}$ point for 4 or 5 days, and no points were given for less than 4 days of proper execution of the behavior in question.

Ratings of exercise and uncontrolled eating were based on actual frequency counts obtained from graphs. For exercise this consisted of a proportion of the total caloric expenditure expected during the program to the number of calories actually performed. Subjects who ate uncontrollably once in a given week received 1 point, twice earned $\frac{1}{2}$ point, and more than twice in a given week earned no points.

These program adherence scores varied from 1.0 to .0. The actual mean scores ranged from a low of .57 for uncontrolled eating to a high of .94 for eating one portion, with the mean for all program behaviors together being .82. It is clear that within the limitations of these self-report data, subjects performed the required program behaviors at or near an optimal level.

Because the distributions of adherence were positively skewed, an arc sine transformation was used to achieve more variability. The transformed measures of program behavior adherence were correlated with weight loss at Week 10, and only uncontrolled eating was significantly correlated with weight change ($r = .29, p < .05$).

With the exception of exercise and self-reinforcement, which were administered to only half of the subjects, the combined relationship of the program adherence measures to weight change was evaluated in a multiple regression analysis. This multiple correlation accounted for an insignificant amount of the variation in weight change at Week 10 ($R^2 = .229, R^2 = .042$). In addition, average adherence scores generated for each subject correlated insignificantly with weight change.

Discussion

Our subjects lost an average of 10.7 pounds (4.8 kg) at program termination, which represents a significant reduction over their initial weight. These results are consistent with previous reports on behavioral approaches to weight reduction, which indicate that subjects lose an average of 1 pound (.45 kg) per week (Leon, 1976; Stunkard, 1975).

Even though these data are noteworthy, other weight reduction regimens such as anorectic drugs and special diets also produce weight loss during their implementation. In most cases, however, once the drug or diet is relinquished, weight gain ensues (Chlouverakis, 1975). So, although behavioral approaches are similar to other forms of weight reduction in demonstrating weight loss during active treatment, whether they foster maintenance of weight loss following program termination is another question.

Maintenance of Weight Loss

To evaluate maintenance, we assessed our results with both 3-month and 1-year follow-ups. The data at the 3-month follow-up show the efficacy of all treatment conditions. Regardless of whether a subject was exposed to the basic program, exercise, contingency management, or some combination thereof, weight loss, on the average, exceeded that at program termination. Of the four conditions, only the *P* group displayed a slight increase (.4 pounds (.2 kg)) over this 3-month period. More importantly, at the 1-year follow-up, treatment gains were maintained in our PEC, PE, and PC groups.

In the studies surveyed by Leon (1976), only four are reported with at least a 6-month follow-up. In Stuart's (1967) report, a mean weight loss at 1 year of 37.7 pounds (17.1 kg) was observed for eight subjects. Comparison with our data is difficult due to the variability in subject sessions (16-41) and the scheduling of maintenance sessions as needed. Also, Mahoney (1974) reported that subjects in an eating habit change group lost 8.3 pounds (3.8 kg) after an 8-week program. At the 1-year follow-up, 70% of these subjects maintained their lower weights or continued to lose. More recently, McReynolds, Lutz, Paulsen, and Kohrs (1976) reported over 17 pounds (7.7 kg) lost after a 15-week program, 19 pounds (8.6 kg) at 3 months, and 17 pounds (7.7 kg) at a 6-month follow-up. Subjects in our PEC, PE, and PC groups performed similarly to those in the McReynolds et al. and Mahoney studies both in terms of the maintenance of weight loss and the percent of subjects maintaining lower weights. Seventy percent of our subjects

(PEC, PE, PC) maintained their lower weights or continued to lose at 3 months and 65% at the 1-year follow-up.

Even though it is clear that most subjects lose weight when involved in the program and many are able to keep it off for extended periods, their ability to *continue losing weight* after program termination is more questionable. Twenty-nine percent of our subjects lost as much as 5 pounds (2.3 kg) more after program termination, whereas only 20% of them were able to lose as much as 10 pounds (4.5 kg) more. Thus, although our subjects were able to maintain their weight loss, less than one third lost a substantial amount of additional weight. Since they averaged approximately 40% overweight, few of them reached their desirable weights over the 1-year time period. In terms of a decrease in percent overweight, the PE group was 12.8% ($p < .001$), the PEC was 9.5% ($p < .01$), the PC was 9.1% ($p < .01$), and the P group was a non-significant 3.8% below the initial values at the 1-year follow-up. Perhaps continued weight loss is best effected through continued in-therapy treatment, booster sessions à la Stuart, or a gradual phasing out of therapy with concomitant buildup of environmental supports.

Generalized Use of Behavioral Programs

Given the overall weight loss and the characteristics of our population (age, socioeconomic status, percent overweight), it does appear that we are able to generalize findings to a nonstudent population. Our subjects ranged in age from 16 to 62 years of age with a mean of 31.5 years. We observed no significant relationship between weight change and age or socioeconomic status, although admittedly the socioeconomic status range was limited.

Exercise

A major focus of this experiment was to evaluate the influence of exercise. A potential problem was our ability to convince subjects of varying ages and degrees of overweight to gradually increase their energy output from 150 to 400 kcalories above their basal levels. Data taken from exercise charts indicated that

the subjects averaged 91% of the required energy output. One might expect older and more obese subjects to have a more difficult time exercising, yet no such relationship was found. Many subjects commented that the gradual increase in exercise, starting at a low 150 kcalories per day, helped them get their previously inactive bodies in shape at a pace that was not too painful. In addition, as the exercise requirement was gradually increased, it gave them time to build up environmental supports for such activity (buying bikes, joining formal exercise programs, etc.). Graphing was viewed as helpful in visualizing progress over time. Also, therapists urged group members to form exercise cohorts among themselves. Finally, on two occasions, therapist models demonstrated appropriate exercises and levels of exertion.

Exercise Versus Contingency Management

This study also provided a contrast of exercise with contingency management whose efficacy has been previously demonstrated (Bellack, 1976; Mahoney, 1974). As previously noted, the main effects of exercise, contingency management, and the interactions were not significant. However, there was a tendency toward an effect for exercise at the 1-year follow-up with mean weight losses of 12.25 and 14.6 pounds (5.5 and 6.6 kg) for those exposed to contingency management and exercise, respectively. In general, this influence of exercise is consistent with the findings of Harris and Hallbauer (1973) and supports its further implementation in weight reduction programs.

Program Adherence

As previously noted, data on program adherence is sparse and limited to self-report questionnaires. Wollersheim (1970) reported differences in eating behavior but did not relate these data to weight loss. However, Mahoney (1974) did show a significant correlation between weight loss and the ability to eliminate inappropriate eating habits. Also, Bellack et al. (1974) found that one of nine program behaviors was significantly related to weight loss.

In contrast to these questionnaire data, we rated program adherence during weekly structured interviews. These data indicated that our subjects engaged in the required behaviors at a reasonably high level. However, as noted, of the 10 behaviors, only uncontrolled eating was significantly related to weight loss, and the combined influence of all behaviors accounted for a small and insignificant proportion of the weight loss variance.

These data on program adherence are somewhat perplexing. The combined effect of our scoring procedure and the response of the subjects provided little variation, which was corrected via transformed scores. Whether this transformation sufficiently enlarged the variation to achieve a significant relationship is mere speculation. If so, the data appear to undermine the presumed influence of the behavioral approach to weight loss. If subjects apply behavioral strategies and change their eating and activity patterns but these variables do not correlate with weight loss, then other so-called nonspecific factors may be operative.

Behavioral Interventions: A Social Influence Process?

The changes in eating and activity patterns may operate within a broader social influence framework to mediate weight loss. In this study, social influence variables were adequately controlled within the therapeutic context. Wollersheim's (1970) comparison of behavioral (focal), nonspecific, and social pressure groups also controlled for the influence of extraneous social variables during "in-therapy" time. However, a characteristic of behavioral approaches is the extremely large amount of "out-of-therapy" time requested of subjects. They record, graph, count, and monitor ad infinitum. These tasks require a great deal of time and activity in the change process.

This effort is unique to behavioral approaches and in marked contrast to other weight reduction regimens. For example, the performance of highly visible "out-of-session" behaviors (e.g., exercise) may increase social reinforcement for adherence well above that available for other less obvious weight reduction regimens such as drugs or diets. It may

well be that any procedure with a credible rationale, behavioral or otherwise, that engages subjects and continuously prompts their attention to weight loss will be effective.

Reference Note

1. Johnson, W. G., & Stalonas, P. M. *Review of program adherence in obesity research: The case for missing independent variables*. Unpublished manuscript, University of Mississippi Medical Center, 1977.

References

- Bellack, A. S. A comparison of self-reinforcement and self-monitoring in a weight reduction program. *Behavior Therapy*, 1976, 7, 68-75.
- Bellack, A. S., Rozensky, R., & Schwartz, J. A comparison of two forms of self-monitoring in a behavioral weight reduction program. *Behavior Therapy*, 1974, 5, 523-530.
- Björntorp, P. Exercise in the treatment of obesity. In M. J. Albrink (Ed.), *Clinics in endocrinology and metabolism*. Philadelphia, Pa.: Saunders, 1976.
- Chlouverakis, C. Dietary and medical treatments of obesity: An evaluative review. *Addictive Behaviors*, 1975, 1, 3-21.
- Feinstein, A. R. The measurement of success in weight reduction: An analysis of methods and a new index. *Journal of Chronic Disease*, 1959, 10, 439-456.
- Hall, S. M., Hall, R. G., Hanson, R. W., & Borden, B. L. Permanence of two self-managed treatments of overweight in university and community populations. *Journal of Consulting and Clinical Psychology*, 1974, 42, 781-786.
- Harris, M. B., & Hallbauer, E. S. Self-directed weight control through eating and exercise. *Behaviour Research and Therapy*, 1973, 11, 523-529.
- Hollingshead, A. B., & Redlich, F. C. *Social class and mental illness*. New York: Wiley, 1958.
- Horan, J. J., Baker, S. B., Hoffman, A. M., & Shute, R. E. Weight loss through variations in the covert control paradigm. *Journal of Consulting and Clinical Psychology*, 1975, 43, 68-72.
- Horton, E. S. The role of exercise in the prevention and treatment of obesity. In G. A. Bray (Ed.), *Obesity in perspective*. Washington, D.C.: U.S. Government Printing Office, 1974.
- Jeffrey, D. B. A comparison of the effects of external control and self-control on the modification and maintenance of weight. *Journal of Abnormal Psychology*, 1974, 83, 404-411.
- Johnson, W. G., & Stalonas, P. M. *Permanent weight loss: The behavioral alternative*. Kalamazoo, Mich.: Behaviordelia, in press.
- Leon, G. R. Current directions in the treatment of obesity. *Psychological Bulletin*, 1976, 83, 557-578.
- Mahoney, M. J. Self-reward and self-monitoring techniques for weight control. *Behavior Therapy*, 1974, 5, 48-57.

- Mahoney, M. J. Fat fiction. *Behavior Therapy*, 1975, 6, 416-418.
- Mahoney, M. J., Moura, N. G. M., Wade, T. C. Relative efficacy of self-reward, self-punishment, and self-monitoring techniques for weight loss. *Journal of Consulting and Clinical Psychology*, 1973, 40, 404-407.
- Mayer, J. *Overweight: Causes, cost and control*. Englewood Cliffs, N.J.: Prentice-Hall, 1968.
- McReynolds, W. T., Lutz, R. N., Paulsen, B. K., & Kohrs, M. B. Weight loss from two behavior modification procedures with nutritionists as therapists. *Behavior Therapy*, 1976, 7, 283-291.
- Metropolitan Life Insurance Company. New weight standards for men and women. *Statistical Bulletin*, 1959, 40, 1-4.
- Stuart, R. B. Behavioral control of overeating. *Behaviour Research and Therapy*, 1967, 5, 357-365.
- Stunkard, A. Presidential address—1974: From explanation to action in psychosomatic medicine: The case for obesity. *Psychosomatic Medicine*, 1975, 37, 195-236.
- Wollersheim, J. P. The effectiveness of group therapy based upon learning principles in the treatment of overweight women. *Journal of Abnormal Psychology*, 1970, 76, 462-474.

Received March 21, 1977 ■

A Longitudinal Study of the Personality Correlates of Marijuana Use

Edwin J. Kay, Arthur Lyons, William Newman,
Donald Mankin, and Roger C. Loeb
Lehigh University

Two hundred and fifty-one male students completed the California Psychological Inventory, the Adjective Check List, and a drug-use questionnaire in the fall of their freshman year and in each of one, two, or three succeeding springs. Three prevailing patterns of drug use were identified. *Continuous nonusers* never disclosed marijuana use; *switched nonusers* did not disclose marijuana use initially but did so on a later questionnaire, and *users* disclosed marijuana use both initially and later. Enduring differences between users and continuous nonusers were found. The switched nonusers generally had scores between those of the users and continuous nonusers. On several scales, switched nonusers were similar to users both before and after their use of marijuana. It is concluded that marijuana use, both present and future, can be predicted by a certain pattern of reported personality characteristics.

A number of investigators have compared users and nonusers of marijuana on a variety of psychosocial and personality measures. Several (Brill & Christie, 1974; McAree, Steffenhagen, & Zheutlin, 1972; Richek, Angle, McAdams, & D'Angelo, 1975) have found few or no significant differences between the two groups. In other cases (Grossman, Goldstein, & Eisenman, 1974; Zinberg & Weil, 1970), group differences have been found only when marijuana use was chronic. However, many researchers have reported significant differences between users and nonusers (Cunningham, Cunningham, & English, 1974; Fisher, 1974; Graham & Cross, 1975; Hogan, Mankin, Conway, & Fox, 1970; Jessor, Jessor, & Finney, 1973; McLaughlin, 1974; Simon, Primavera, Simon, & Orndoff, 1974). The great majority

of the studies cited above have used data collected at one testing only and thus do not show whether the personality differences sometimes found were antecedent or consequent to marijuana use. In addition, they could provide no information on the consistency of the measures over time and on the characteristics of individuals who change their patterns of drug use.

Two studies (Brill & Christie, 1974; Jessor et al., 1973) did contain a longitudinal design. Brill and Christie found only slight differences between users and nonusers and therefore did not report longitudinal data. In the Jessor et al. (1973) study, high school students were classified into one of three groups: nonusers of marijuana initially and a year later (NU-NU); nonusers of marijuana initially who became users of marijuana 1 year later (NU-U); and users initially and 1 year later (U). On a number of personality measures, the NU-U subjects initially fell between the NU-NU and U groups, differing significantly from NU-NU subjects on a number of dimensions. For the NU-NU subjects there were no significant changes over time. In contrast, for the NU-U subjects there were significant changes in which they became more similar to the U

This research was supported by National Institute of Mental Health Grant R01DA0054. We thank the anonymous reviewers for their valuable suggestions.

Arthur Lyons is now at Moravian College, Bethlehem, Pennsylvania.

Donald Mankin is now at the University of Houston at Clear Lake City, Houston, Texas.

Requests for reprints should be sent to Edwin J. Kay, Department of Psychology, Chandler-Ullmann Hall #17, Lehigh University, Bethlehem, Pennsylvania 18015.

subjects on numerous measures. In summary, high school respondents who switched from nonuse to use of marijuana were initially between nonusers and users, and they became more like users over time. Jessor et al. also included a sample of college students in their study, but unfortunately they did not collect longitudinal data on these subjects.

The present study is a modified replication of Hogan et al. (1970) in that the subjects were from Lehigh University and some of the same personality measures were used. In that study, users and nonusers differed on such measures as socialization, flexibility, and empathy. The addition of a longitudinal design (similar to that of Jessor et al., 1973) makes it possible to assess the consistency of the reported differences in characteristics of users and nonusers over time and to check for personality characteristics and changes in nonusers who became users. Concurrent validity could also be evaluated, since more than one personality measure was used.

Method

Subjects

Each fall during the years 1971, 1972, 1973, 200 randomly chosen male freshmen entering Lehigh University were solicited for voluntary participation. The various groups of students were then contacted for follow-up studies in the spring following initial contact and for successive springs up to and including the spring of 1974. Thus, freshmen entering in the fall of 1971 were contacted then and in the springs of 1972, 1973, and 1974; freshmen entering in the fall of 1972 were contacted then and in the springs of 1973 and 1974; freshmen entering in the fall of 1973 were contacted then and in the spring of 1974. Initially, 130 (65%) of the 200 students who were contacted in the fall of 1971 agreed to be in the study; the corresponding numbers of freshmen entering in the fall of 1972 and 1973 were 124 (62%) and 112 (56%), respectively. The numbers of students participating in the initial testing and all follow-up testing were 68, 85, and 98 for the freshmen entering in the fall of 1971, 1972, and 1973, respectively. Each subject was paid for each testing session in which he participated.

The problem of subject loss is attributable to several factors. First, only students remaining at Lehigh for the duration of the research (1-3 years) were included in the final samples. Second, among students remaining at Lehigh, some could not be located during the testing weeks. And third, some students who were contacted were unwilling or unable to participate in all testing sessions. This resulted in the following percentages of original volunteers completing all testing sessions:

52% of those entering in 1971, 69% entering in 1972, and 87% entering in 1973. These figures are similar to other longitudinal research in the area (Brill & Christie, 1974; Jessor et al., 1973).

Procedure

For each test session, the subjects completed three instruments: a drug questionnaire (including demographic as well as drug use information); the Adjective Check List (ACL; Gough & Heilbrun, 1965); and the California Psychological Inventory (CPI; Gough, 1964). Subjects reported varying levels of marijuana use, and a small minority reported the use of other illegal drugs. The clearest distinction among subjects was whether or not any marijuana use was disclosed for a given test session.

Results

The analyses reported below are only for those subjects who participated in all of the appropriate testing sessions of the experiment. According to their responses on the various administrations of the drug questionnaire, the subjects were assigned to one of four categories. A subject was categorized as a *continuous non-user* (CNU) if he did not disclose use of marijuana on any administration of the drug questionnaire. A subject was categorized as a *user* (U) if he disclosed use of marijuana on the initial administration of the drug questionnaire. A subject was categorized as a *switched nonuser* (SNU) if he did not disclose marijuana use at the initial administration of the drug questionnaire but did disclose such use on some subsequent administration of the drug questionnaire. No subjects were found in the fourth possible category, *switched user* (disclosure of marijuana use on the initial administration of the drug questionnaire but denial of marijuana use on later questionnaires).

Replication of Hogan et al.

Our first aim was to see if our data replicated the study of Hogan et al. (1970), who administered the CPI to comparable subjects. Table 1 shows the raw score means on the CPI scales for U and CNU subjects for various administrations of the test. The data from the fall and first spring are for subjects who participated in the experiment for 1, 2, or 3 years; the data from the second spring are for those subjects who participated in the experiment for 2 or 3

Table 1

Mean Responses on the California Psychological Inventory Scales for User (U) and Continuous Nonuser (CNU) Subjects

Scale	Time of administration							
	Fall		1st spring		2nd spring		3rd spring	
	U (78)	CNU (115)	U (78)	CNU (115)	U (38)	CNU (68)	U (12)	CNU (27)
Dominance	25.3	24.2	26.4	26.0	26.3	27.6	24.9	28.4
Capacity for Status	17.6	17.2	18.7	17.4*	18.8	18.7	20.4	18.5
Sociability	23.9	22.1**	24.1	22.6	24.3	24.0	25.5	24.7
Social Presence	37.4	33.4***	39.5	35.6***	39.0	36.1*	41.6	36.2*
Self-acceptance	21.8	20.6*	21.9	21.5	21.6	21.4	21.9	21.8
Well-Being	32.6	33.4	32.8	33.9	31.7	33.3	33.2	33.0
Responsibility	25.0	28.1***	25.3	28.1***	25.3	28.6***	24.7	28.8*
Socialization	33.7	38.0***	32.8	37.7***	31.7	36.8***	30.2	37.1***
Self-control	23.0	27.3***	22.5	26.4***	22.0	27.1***	24.1	27.0
Tolerance	18.9	19.4	20.4	20.1	20.2	20.7	22.4	20.5
Good Impression	12.9	15.2***	13.6	14.3	13.6	15.6	13.8	15.2
Communality	24.1	24.3	23.1	24.7***	23.0	23.6	23.8	24.0
Achievement via Conformance	22.8	26.1***	23.4	26.5***	23.2	26.9***	24.1	26.1
Achievement via Independence	19.1	19.4	19.5	19.8	20.3	20.2	22.3	20.2
Intellectual Efficiency	35.6	36.6	37.0	37.1	36.7	37.9	39.1	38.1
Psychological-mindedness	11.7	11.4	11.6	11.8	10.8	11.9	12.2	12.0
Flexibility	12.7	9.8***	12.7	10.2***	13.8	10.7***	14.3	10.9**
Femininity	16.7	17.5	15.7	16.7	16.9	17.0	16.2	17.6

Note. Asterisks indicate significant *t* tests between U and CNU means. Numbers in parentheses are *ns*.

* $p < .05$.

** $p < .025$.

*** $p < .01$.

Table 2

Mean Scores of User (U), Switched Nonuser (SNU), and Continuous Nonuser (CNU) Subjects for Selected California Psychological Inventory Scales

Scale	Time of administration											
	Fall			1st spring			2nd spring			3rd spring		
	U	SNU	CNU	U	SNU	CNU	U	SNU	CNU	U	SNU	CNU
	(78)	(58)	(115)	(78)	(58)	(115)	(38)	(47)	(68)	(12)	(29)	(27)
Social Presence	37.4	37.3	33.4**	39.5	36.7	35.6	39.0	38.0	36.1	41.6	39.4	36.2
Responsibility	25.0*	27.5	28.1	25.3	26.6	28.1	25.3	26.7	28.1	24.7	26.3	28.8
Socialization	33.7**	37.2	38.0	32.8*	35.6	37.7	31.7	34.4	36.8	30.2**	36.5	37.1
Self-control	23.0	26.0	27.3	22.5	25.2	26.4	22.0	25.0	27.1	24.1	26.5	27.0
Achievement via												
Conformance	22.8**	26.0	26.1	23.4	24.6	26.5	23.2	24.9	26.9	24.1	27.2	26.1
Flexibility	12.7	11.1	9.8	12.7	11.6	10.2*	13.8	12.9	10.7	14.3	14.0	10.9*

Note. Asterisks next to scores indicate significant *t* tests when compared to SNU means. Numbers in parentheses are *ns*.

* $p < .05$.

** $p < .01$.

years; and the data from the third spring are for those subjects who participated in the experiment for 3 years (i.e., those who entered Lehigh in 1971). If the four drug use categories of Hogan et al. (frequent users, occasional users, nonusers, and principled nonusers) are collapsed into two categories (users and non-users), then their Table 3 data can be compared with our fall data. The comparison indicates that the results of our initial fall testing closely replicated Hogan et al. Furthermore, the pattern of results endured over time as assessed by our three retests. Differences that both replicated Hogan et al. and endured over time occurred on six scales. Compared to CNU subjects, U subjects were significantly higher in Social Presence and Flexibility and lower on Responsibility, Socialization, Self-control, and Achievement via Conformance.

In Table 2, we present the mean CPI scores for all three categories of subjects on the six scales that consistently differentiated U and CNU subjects over time. SNU subjects generally fell between scores of U and CNU subjects. They were consistently similar to U subjects on the Social Presence scale and similar to CNU subjects in the Socialization, Self-control, and Achievement via Conformance scales. SNU subjects shifted on only one CPI scale; over time they came to resemble the U subjects on Flexibility.

Adjective Check List

The mean responses on the ACL by U and CNU subjects for various administrations of the test are presented in Table 3. Again the *t* test was used to compare the scale means for

Table 3
Mean Responses on the Adjective Check List Scales for User (U) and Continuous Nonuser (CNU) Subjects

Scale	Time of administration							
	Fall		1st spring		2nd spring		3rd spring	
	U (78)	CNU (115)	U (78)	CNU (115)	U (38)	CNU (68)	U (12)	CNU (27)
No. adjectives checked	49.4	49.4	50.8	51.2	51.6	50.8	50.7	55.1
Defensiveness	48.8	50.9	49.1	51.0	48.5	51.0	46.9	50.5
Favorable adjectives checked	47.0	48.4	47.6	49.8	46.1	50.7*	45.5	50.8
Unfavorable adjectives checked	51.7	50.2	51.6	50.4	51.6	49.6	51.7	48.1
Self-confidence	46.2	44.4	46.7	45.3	46.1	46.4	40.5	44.6
Self-control	44.9	50.8**	44.0	51.6**	45.5	50.4**	45.2	50.9
Libality	54.9	49.6**	56.6	49.9**	57.6	51.3**	58.5	49.7**
Personal Adjustment	46.6	49.0*	45.8	49.9**	45.8	49.7*	42.5	52.3**
Achievement	48.2	51.9**	49.0	52.4*	47.3	52.2*	44.5	50.6
Dominance	48.8	48.8	50.6	48.9	48.4	49.5	45.7	48.7
Endurance	47.2	53.0**	47.3	53.0**	46.6	52.5**	45.9	51.3
Order	46.6	53.5**	46.4	54.0**	46.6	53.8**	44.2	53.7**
Introception	50.3	51.8	50.4	51.7	50.6	54.6	49.8	56.6*
Nurturance	49.1	49.2	48.6	50.4	49.1	51.1	50.3	53.1
Affiliation	48.4	48.1	48.7	50.1	47.7	49.5	47.8	49.0
Heterosexuality	53.1	45.6**	54.0	46.9**	51.9	50.4	53.6	50.4
Exhibition	50.0	45.1**	51.7	45.7**	51.4	47.6	50.5	44.6
Autonomy	51.8	47.3**	53.7	47.6**	53.4	48.1**	50.8	46.7
Aggression	49.6	47.0	51.5	46.5**	49.8	46.8	48.4	45.3
Change	51.6	45.3**	53.6	46.5**	52.6	46.7**	54.2	44.4**
Succorance	48.3	49.3	46.5	48.2	47.9	47.7	50.1	46.1
Abasement	49.1	51.9	47.2	50.8**	48.0	50.6	51.5	51.0
Deference	47.2	52.5**	45.3	51.9**	47.2	52.2*	49.2	53.6
Counseling Readiness	50.2	53.5	49.0	52.3	50.1	51.4	53.5	53.4

Note. Asterisks indicate significant *t* tests between U and CNU means. Numbers in parentheses are *ns*.

* $p < .05$.

** $p < .01$.

U and CNU subjects. On the initial administration, U and CNU subjects differed significantly at the .01 level on 10 of the 25 scales; for one scale the difference was significant at the .05 level. Users were significantly higher than CNU subjects on Liability, Heterosexuality, Exhibition, and Change. CNU subjects were significantly higher than U subjects on Self-control, Personal Adjustment, Achievement, Endurance, Order, Autonomy, and Deference.

In examining Table 3, we see that the original differences on these 11 scales held up consistently over the 3 years of the study. In the spring of the third year, the magnitude of the differences was the same, although some differences were not statistically significant. Only subjects who entered school in the fall of 1971 were measured in the spring of the third year; this drop in sample size substantially reduced the power of the statistical test. Nonetheless, it is impressive that 2½ years after the first test, the direction of the differences remained the same.

The data for the 11 ACL scales for the SNU subjects as compared to U and CNU subjects are presented in Table 4. For the first two test

sessions, with only one exception (Personal Adjustment for first spring), SNU subjects fell between U and CNU subjects on these 11 scales. However, in the last two test sessions, this pattern did not occur. Thus, the differences among the three groups on the 11 scales changed over time. A careful examination of Table 4 reveals a fairly consistent pattern. Over time, many of the ACL scale scores of the SNU subjects become more like those of the U subjects. The shift from similarity with CNU subjects to similarity with U subjects is clear on the following scales: Socialization through Conformity, Liability, Autonomy, and Change. However, on the Order, Heterosexuality, and Exhibition scales, the SNU subjects were consistently similar to U subjects.

Finally, to assess concurrent validity, we hypothesized that the significant differences found between U and CNU subjects would be in the same direction for the CPI and the ACL when the two scales are positively correlated (Gough & Heilbrun, 1965) and in opposite directions when the two scales are negatively correlated. Fifty-eight of the 66 possible cases supported this hypothesis; that is, the CPI and ACL scores were in the predicted direction.

Table 4

Mean Scores of User (U), Switched Nonuser (SNU), and Continuous Nonuser (CNU) Subjects for Selected Adjective Check List Scales

Scale	Time of administration											
	Fall			1st spring			2nd spring			3rd spring		
	U (78)	SNU (58)	CNU (115)	U (78)	SNU (58)	CNU (115)	U (38)	SNU (47)	CNU (68)	U (12)	SNU (29)	CNU (27)
Self-control	44.9**	48.5	50.8	44.0	47.1	51.6	45.5	44.6	50.4*	45.2	45.2	50.9
Liability	54.9**	50.3	49.6	56.6	53.3	49.9	57.6	54.4	51.3	58.5	58.9	49.7
Personal Adjustment	46.4	47.1	49.0	48.8	46.9	49.9	46.8	42.4	49.7*	42.5**	48.2	52.3
Achievement	48.2	50.7	51.9	49.0	50.2	52.4	47.3	48.0	52.2	44.5**	52.7	50.6
Endurance	47.2	50.5	53.0	47.3	49.8	53.0	46.6	46.0	52.5**	45.9	49.1	51.3
Order	46.6	49.1	53.5**	46.4*	50.7	54.0	46.6	46.6	53.8**	44.2	46.7	53.7*
Heterosexuality	53.1	52.1	45.6**	54.0	52.2	46.9**	51.9	52.5	50.4	53.6	57.7	50.4
Exhibition	50.0	48.8	45.1**	51.7	50.3	45.7	51.4	52.9	47.6	50.5	56.9	44.6*
Autonomy	51.8	48.4	47.3	53.7**	49.6	47.6	53.4	54.0	48.1**	50.8	56.2	46.7*
Change	51.6*	46.6	45.3	53.6	48.0	46.5	52.6	48.8	46.7	54.2	54.0	44.4*
Deference	47.2	50.2	52.5	45.3**	49.1	51.9	47.2	45.5	52.2*	49.2	43.2	53.6*

Note. Asterisks next to scores indicate significant *t* tests when compared to SNU means. Numbers in parentheses are *ns*.

* $p < .05$.

** $p < .01$.

The probability of finding this amount of agreement by chance is extremely low ($p < .001$). Furthermore, on the correlations between CPI and ACL scores, 32 of the 58 correlations in agreement with our hypothesis were significant, whereas none of the 8 cases in disagreement with our hypothesis was significant.

Discussion

Comparing our data to those of Hogan et al. (1970), we are impressed that the personality correlates of marijuana use among the reputedly "straight," apathetic, job-oriented college youth of the early 1970s are strikingly similar to the personality correlates associated with marijuana use among the "hippie" activist college youth of the late 1960s. Although the present sample of college males at a relatively small and conservative university is not likely to be representative of American youth, the subjects are probably similar to a large number of college students. This successful replication of Hogan et al. provides support for the findings of both studies and implies that the personality patterns associated with marijuana use and nonuse have not changed significantly over the years.

The consistency between studies and the consistency across measures and across years (at least on the CPI) within our study suggests that certain people with specifiable personality patterns fall into our three categories of continued nonusers, users, and switched nonusers. The distinction between users and nonusers is clearest in three personality clusters. First, nonusers tend to be higher in several indices of what might be labeled *conformity*. On the CPI, nonusers had high scores on three of the group of scales described by Gough as "measures of socialization, maturity, responsibility, and intrapersonal structuring of values" (1964, p. 10). These three scales (Responsibility, Socialization, and Self-control) all reflect a concern with social standards. The low scores of the drug users reflect relative irresponsibility, rebelliousness, and hostility to rules and conventions compared to nonusers. It is interesting to note that even nonusers in our sample scored below Gough's (1964) male norms on two (Responsibility, Self-control) of the scales.

Conformity among nonusers is also reflected on the ACL scales of Deference, Order, and Self-control. On the other hand, the individuality and lack of conformity of the user is evidenced in high ACL scores on Autonomy, Change, Exhibition, and Liability.

The second distinguishing personality cluster is related to the issue of socialization and conformity. It is a set of characteristics that increases the likelihood of success through conventional means. Nonusers are relatively high in Achievement through Conformance on the CPI. This scale indicates efficiency, organization, and industriousness. Similarly, on the ACL, nonusers were higher on Achievement and Endurance (persistence). They also demonstrated a more positive attitude toward their place in society by their superior scores on Personal Adjustment.

The third cluster of traits involves adventure seeking. On the CPI, users evinced more spontaneity (measured by Social Presence) and adventuresomeness (Flexibility). Similarly, users appeared high on adventure (Heterosexuality) and seeking novelty (Change) according to the ACL.

In summary, nonusers appear to be well socialized. They conform to norms, respect authority, strive for traditional goals, and rarely act on impulse. Users show a strikingly different picture. They are nonconforming, independent, adventurous, and spontaneous. This pattern generally supports contemporary myths about college-age drug users and nonusers. It should be remembered that these data are based on self-report measures; that is, they tell us how the subjects perceive, or claim to perceive, themselves. Thus, marijuana users and nonusers may in part be reflecting social expectations. How their self-perceptions relate to others' perceptions or to reality (however it is defined) has not been assessed.

At any rate, that users and nonusers reliably report self-perceptions that sharply differ is both interesting and potentially useful. The usefulness of these data becomes clearer when we examine the third group, the switched nonusers. It is noteworthy that approximately 90% of the SNU subjects had switched to marijuana use by the second spring testing. Thus, the data in the last two testings reflect postmarijuana use personality characteristics.

In general, this group fell between the users and the continuous nonusers on the various personality scales; this outcome replicates the study by Jessor et al. (1973). If we are interested in predicting which nonusers have the greatest chance of becoming users, there are several CPI and ACL scales that appear to be useful. Specifically, if SNU subjects show certain personality characteristics similar to U subjects (and dissimilar to CNU subjects) both before and after their drug use, those characteristics can be used as predictors of future drug use. SNU subjects were consistently similar to U subjects on the Social Presence scale of the CPI and on the Order, Heterosexuality, and Exhibition scales of the ACL. The SNU group can thus be described as consistently similar to the U group in being outgoing, socially self-confident, and spontaneous. One might say that the "extraverted personality" is susceptible to the use of marijuana.

Several studies have reported findings that support the conclusion that marijuana users are extraverts. Hogan et al. (1970) described the marijuana user as high in social poise (though this was offset by an "assertive non-conformity"). According to Brill & Christie (1974), marijuana users reported themselves as higher in the tendency to seek stimulation. Simon et al. (1974) reported users to be low in deference and order. Graham and Cross (1975) found that users value feelings and experience over planning and logic, an interpersonal responsiveness that seems to characterize our switched nonusers.

On the other hand, SNU subjects shifted from similarity with CNU subjects to similarity with U subjects on the Flexibility scale of the CPI, and on the Socialization through Conformity, Lability, Autonomy, and Change scales of the ACL. This second cluster of personality traits involving nonconformity, lack of responsibility, and change is reflected in most of the literature on drug users (e.g., Cunningham et al., 1974; Grossman et al., 1974; Hogan et al., 1970; Simon et al., 1974; Zinberg & Weil, 1970). However, these studies only compared users with nonusers. Thus, the characteristics that SNU subjects shared with U subjects after they started using marijuana appear to be excellent post hoc discriminators;

they are not useful in discriminating nonusers who later become users from nonusers who continue not to use marijuana.

Two final points relate to the personality tests. First, contrary to the claims of some critics of the CPI, this longitudinal study found little evidence of shifts in personality results in conjunction with changes in drug use. Our SNU subjects changed on only one CPI scale—Flexibility. This supports our earlier contention that a certain personality type may be recruitable to marijuana use rather than the typical interpretation that marijuana use results in personality changes.

The second point is that in contrast to the consistency of the CPI scores, some ACL scores did shift among SNU subjects. On the ACL, SNU subjects demonstrated a trend over time toward similarity to users. As the ACL was developed to assess self-concept, this difference may imply that SNU subjects are experiencing and reporting changes in their self-concepts. Following their participation in marijuana use, the SNU subjects developed a self-concept of themselves similar to marijuana users. This change is not reflected in personality trait modifications as measured by the CPI. Perhaps such situational influences as drug use and associated environmental conditions have impact on ACL-type self-concept without resulting in changed perceptions of personality characteristics measured by the CPI.

References

- Brill, N. Q., & Christie, R. L. Marijuana use and psychosocial adaptation. *Archives of General Psychiatry*, 1974, 31, 713-719.
- Cunningham, W. H., Cunningham, I. C. M., & English, W. D. Sociopsychological characteristics of undergraduate marijuana users. *Journal of Genetic Psychology*, 1974, 125, 3-12.
- Fisher, G. Milieu of marijuana use. *International Journal of Social Psychiatry*, 1974, 20, 45-55.
- Gough, H. G. *Manual for the California Psychological Inventory* (Rev. ed.). Palo Alto, Calif.: Consulting Psychologists Press, 1964.
- Gough, H. G., & Heilbrun, A. B., Jr. *The Adjective Check List manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1965.
- Graham, D. L., & Cross, W. C. Values and attitudes of high school drug users. *Journal of Drug Education*, 1975, 5, 97-107.
- Grossman, J. C., Goldstein, R., & Eisenman, R. Undergraduate marijuana and drug use as related to openness of experience. *Psychiatric Quarterly*, 1974, 48, 86-92.

- Hogan, R., Mankin, D., Conway, J., & Fox, S. Personality correlates of undergraduate marijuana use. *Journal of Consulting and Clinical Psychology*, 1970, 38, 58-73.
- Jessor, R., Jessor, S., & Finney, J. A social psychology of marijuana use: Longitudinal studies of high school and college youth. *Journal of Personality and Social Psychology*, 1973, 26, 1-15.
- McAree, C. P., Steffenhagen, R. A., & Zheutlin, L. S. Personality factors and patterns of drug usage in college students. *American Journal of Psychiatry*, 1972, 128, 890-893.
- McLaughlin, J. F. Selected personality characteristics of the moderate marijuana, heavy marijuana, and poly-drug-using marijuana smoker. (Doctoral dissertation, University of Montana, 1973). *Dissertation Abstracts International*, 1974, 34, 3881A. (University Microfilms No. 74-1390).
- Richek, H. G., Angle, J. F., McAdams, W. S., & D'Angelo, J. Personality/mental health correlates of drug use by high school students. *Journal of Nervous and Mental Disease*, 1975, 160, 435-442.
- Simon, W. E., Primavera, L. H., Simon, M. G., & Orndoff, R. K. A comparison of marijuana users and nonusers on a number of personality variables. *Journal of Consulting and Clinical Psychology*, 1974, 42, 917-918.
- Zinberg, N. E., & Weil, A. T. A comparison of marijuana users and nonusers. *Nature*, 1970, 226, 119-123.

Received April 4, 1977 ■

The Child Behavior Profile: I. Boys Aged 6-11

Thomas M. Achenbach

National Institute of Mental Health, Bethesda, Maryland

The goal was to develop a descriptive classification system that could be used to group children for research and clinical purposes, to reflect adaptive competencies as well as behavior problems, and to facilitate quantitative assessment of behavioral change. The system is embodied in a series of Child Behavior Profiles that are standardized separately for children of each sex at ages 4-5, 6-11, and 12-16. The profiles are scored from the Child Behavior Checklist (CBCL), which was designed to obtain parents' reports of their children's competencies and problems in standardized format. This article reports standardization of the profile for boys aged 6-11. Factor analysis of the CBCLs of 450 disturbed boys yielded nine behavior problem scales labeled *Schizoid*, *Depressed*, *Uncommunicative*, *Obsessive-Compulsive*, *Somatic Complaints*, *Social Withdrawal*, *Hyperactive*, *Aggressive*, and *Delinquent*. The first five problem scales load on a second-order factor labeled *Internalizing*, and the last three load on a factor labeled *Externalizing*. Three social competence scales entitled *Activities*, *Social*, and *School* were also constructed from the CBCL. Norms are based on a normal sample of 300. Comparison of disturbed and normal boys showed differences ($p < .001$) on all behavior problem and social competence scores. Eight-day test-retest correlations averaged .89, whereas interparent correlations averaged .74. Computerized and hand-scored versions of the profile can be used to display item and scale scores for individual boys.

One of the greatest handicaps to research and communication on child psychopathology has been the lack of a standardized, objective, and reliable way of describing and classifying behavior disorders. Not until the 1968 edition of the American Psychiatric Association's *Diagnostic and Statistical Manual* (DSM; APA, 1968) was the need for a differentiated classification of children's disorders even recognized in the official nomenclature. Prior to the 1968 edition, the only childhood disorders recognized

by the DSM were adjustment reaction of childhood and childhood schizophrenia.

The author is indebted to the staffs of the many clinical agencies that have contributed to this work, as well as to Sue Fleisher for her unflagging efforts in all phases of the research, Melvin Lewis for his collaboration in developing the behavior problem component of the CBCL, and Brian McLaughlin and Craig Edelbrock for programming the Child Behavior Profile.

Requests for reprints, for copies of the CBCL, and for computerized and hand-scored versions of the Child Behavior Profile should be sent to Thomas M. Achenbach, Laboratory of Developmental Psychology, Building 15K, National Institute of Mental Health, 9000 Rockville Pike, Bethesda, Maryland 20014.

Although provisional guides to classification such as the DSM may be needed during the early stages of a field's development, the inadequacy of the DSM has inspired numerous attempts to evolve alternative methods of classification for childhood disorders. The most common approach has been to factor analyze checklists of behavior problems. However, the diversity of checklists, subject samples, sources of data, and methods of analysis has led to almost equally great diversity of results (Achenbach & Edelbrock, in press). One of the recurrent differences among studies is whether they yield a small number of broad-band factors or a larger number of narrow-band factors. The number and breadth of the factors has been determined largely by the number of items used and the methods of analysis. At one extreme, Quay and Peterson's (1967) 55-item Behavior Problem Checklist typically yields two broad-band factors labeled *Conduct Problem* and *Personality Problem*, although

two smaller factors labeled *Inadequacy-Immaturity* and *Socialized Delinquency* have also been found (cf. Quay, 1972). At the opposite extreme, Baker and Dreger (1973) have derived 30 factors from a checklist of 274 items.

Findings of a few broad factors and numerous narrower factors are not necessarily contradictory, as second-order analyses have shown narrow-band factors to be subsumed by broad groupings like the Conduct Problem and Personality Problem factors (Achenbach, 1966; Miller, 1967). Other analyses by Achenbach (1966) have shown that the two broad-band groupings—which he labeled *Externalizing* and *Internalizing*—were replicated in child psychiatric samples differing in age, sex, and socioeconomic status, but that the narrow-band factors differed among various subgroups. It thus appears that the broad-band factors represent general behavior patterns but may mask syndromes that vary with such characteristics as sex and developmental level.

The value of any classification system depends on the function it is to serve. In our present state of ignorance about etiology, prognosis, and appropriate treatment, the worth of a system for describing and classifying child psychopathology can be measured in terms of the following criteria: (a) It should provide a description of behavior in a standardized format that is useful to clinicians and researchers alike. (b) It should be differentiated enough to include narrow-band syndromes peculiar to particular subgroups. (c) It should not rest on clinical inferences by professionals, as few children in need of help receive adequate professional attention. (d) It should reflect children's positive adaptive competencies as well as their maladaptive characteristics. (e) It should enable us to group children for purposes of research on etiology, epidemiology, and treatment effectiveness. (f) It should facilitate quantitative assessment of behavioral change in order to evaluate prognosis under various conditions.

This article reports efforts to develop a descriptive classification system that will fulfill the six criteria just enumerated. Separate analyses are performed for children divided according to age and sex in order to detect patterns that may be peculiar to particular subgroups. Data are obtained with the Child

Behavior Checklist (CBCL), which comprises not only a diverse array of behavior problems but also items reflecting adaptive competencies, including participation in various activities, social relationships, and school success. These items form three social competence scales on which children are scored in relation to norms for their age and sex. The CBCL, which takes about 17 minutes to complete, is designed to be filled out by parents or parent surrogates, because they typically have a more comprehensive picture of their children's problems and competencies than do any other possible informants. Furthermore, parents' views and biases are pivotal in determining whether clinical services are obtained and which treatment options are implemented; they are also important in determining the long-term prognoses.

Data obtained with the CBCL are entered on the Child Behavior Profile, which displays the items reported by parents as well as the child's standing on narrow- and broad-band syndromes. Using either a computerized or a hand-scored version of the profile, the clinician or researcher can obtain an overview of the specific behavior reported by the parent, how the child's problems and competencies cluster, and how the child compares with normal children of similar age and sex. The profile approach preserves more information than does classification into mutually exclusive categories according to individual syndromes, and the profiles themselves can be used as a basis for multidimensional classification. The profile described here was constructed from data on normal and clinical samples of boys 6–11 years old. Subsequent articles will report the profiles for boys 4–5 and 12–16, and girls 4–5, 6–11, and 12–16.

Method

CBCL Behavior Problem Items

Development of the CBCL began with the behavior problem checklist that was constructed by Achenbach (1966) from a survey of existing literature and case histories of 1,000 child psychiatric patients. The original checklist was designed to be filled out from case history data by raters using present-absent response alternatives. It was adapted for parents' use by simplifying the wording, expanding the present-absent alternatives

to a 0-1-2 scale, and adding new items in consultation with clinicians. Pilot editions were further revised on the basis of item analyses and feedback from parents, clinicians, and paraprofessionals.

The current edition comprises 118 behavior problem items to which the parent responds by circling a 0, 1, or 2 according to the following instructions:

Below is a list of items that describe children. For each item that describes your child *now* or *within the past 12 months*, please circle the 2 if the item is *very true* or *often true* of your child. Circle the 1 if the item is *somewhat* or *sometimes true* of your child. If the item is *not true* of your child, circle the 0.

The items are intended to provide broad but non-redundant coverage of behavioral problems that can be rated with a minimum of inference. The parent is requested to write in descriptions of behaviors for items that might otherwise be ambiguous. For example, Item 28 is: Eats or drinks things that are not food (describe) _____. Parents' descriptions make it possible to discriminate between those who are concerned about their child's consumption of junk foods and those whose child is eating dirt, paint, and so on. Only nonfood substances such as the latter are scored on the Child Behavior Profile. In addition to the 118 items, spaces are provided for parents to write in unlisted physical problems having no known medical cause and any other problems that are not listed.

CBCL Social Competence Items

Following a survey of the meager existing literature on social competence indices for children, descriptions of positive behavioral characteristics were piloted in various formats with parents. It was found that items paralleling the behavior problems but describing positive characteristics inevitably sounded like lists of boy scout virtues with a strong social desirability component. Most parents endorsed all such items as describing their child. On the other hand, items of the type used on the Vineland Social Maturity Scale (Doll, 1965) failed to discriminate among children of normal intelligence. The items found to be most successful and ultimately selected for the CBCL comprise scales of involvement and attainment in the three areas described below.

Activities scale. This scale consists of scores for the amount and quality of the child's participation in (a) sports; (b) nonsports hobbies, activities, and games; and (c) jobs and chores. The parent is first asked to report the participant sports (up to three) that the child likes best. Beside each sport, the parent is asked to check boxes indicating how much time—compared to other children of the same age—the child spends in the sport and boxes indicating how well—compared to other children of the same age—the child does in each one. The alternatives include *don't know*, *below average*, *average*, and *above average*. A second item, with similar response alternatives, is provided for reporting the child's favorite hobbies, activities, and games, other than sports. On the third item of the Activities scale,

the parent is to list the child's jobs or chores (up to three). Beside each entry the parent is to check boxes indicating how well the child carries it out, compared to other children of the same age. The response alternatives are like those of the preceding items.

After trying other response formats, such as requesting parents to report the actual frequency of each activity, the present format was chosen for three reasons: (a) The significance of the frequency of participation varies greatly with the particular activity and environment. (e.g., Opportunities for riding bicycles are generally more frequent than for skiing, but both depend on the season of the year and the locality.) (b) To make the parent's task as easy as possible, we wished to use a format of maximum simplicity and generality. (c) Recognizing that we are obtaining the parent's perception of the child, we wished to maximize the power of the CBCL to discriminate children for whom parents could report at least some evidence of social competence from those for whom parents could report nothing positive.

The scoring system for the Activities scale allocates 0-2 points for number of sports, with 0 being assigned if one or no sport is listed, 1 point for two sports, and 2 points for three sports; 0-2 points for the mean score obtained for sports participation and skill, with 0 assigned for each response of *below average*, 1 for each response of *average*, 2 for each response of *above average*, and *don't know* responses omitted from scoring; 0-2 points for number of activities; 0-2 points for average amount and quality of participation in activities; 0-2 points for number of jobs and chores; and 0-2 points for average quality of job performance. The latter four scores follow the same scoring principles as outlined for sports.

The reason for assigning a score of 0 to a report of either zero or one sport, activity, or job is that so few parents of children in clinical samples reported no sports, other activities, or jobs that the difference between none and one did not appear worth recognizing in the scoring. However, a child who has none can receive only a 0 for participation and skill in that area, whereas a child who has one can receive up to 2 points for participation and skill for that one entry. The six scores for the Activities scale are added together to provide a summary score that can range from 0 to 12, and this is entered with its *T* score on the Child Behavior Profile, as explained later.

Social scale. The Social scale consists of scores for (a) the child's membership and participation in organizations; (b) number of friends and contacts with them; and (c) behavior with others and alone. On the first item, the parent is asked to list (up to three) organizations, clubs, teams, or groups the child belongs to and to indicate how active the child is in each, compared to other children of the same age. The number of organizations and amount of participation in each are scored 0-2 in the same fashion as items on the Activities scale. On the next item, the parent is to indicate how many close friends the child has, with the response alternatives being *none*, 1, 2 or 3, and 4 or more. *None* and 1 are both scored 0, 2 or 3 is scored 1, and 4 or more is scored 2. The parent is also asked to indicate how

many times a week the child does things with his/her friends. The responses—less than 1, 1 or 2, and 3 or more—are scored 0, 1, and 2, respectively.

The third item of the Social scale asks: "Compared to other children of his/her age, how well does your child: Get along with his/her brothers and sisters? Get along with other children? Behave with his/her parents? Play and work by himself/herself?" The response alternatives are *worse*, *about the same*, and *better*, and they are scored 0, 1, and 2, respectively. Responses to the first three questions are averaged to provide a score for behavior with others, whereas the response to the last question provides a score for independent behavior. The Social scale score is the sum of the six scores just described, each of which can range from 0 to 2, for a possible total of 12.

School scale. The School scale consists of scores for (a) the average of the child's performance in academic subjects; (b) placement in a regular or special class; (c) being promoted regularly or held back; and (d) the presence or absence of school problems. For academic performance, the alternatives are *failing*, *below average*, *average*, and *above average* for reading, writing, arithmetic, spelling, and/or other subjects. The response alternatives are scored 0, 1, 2, and 3, respectively, and are averaged to provide a score ranging from 0 to 3.

The parent is next asked to indicate whether the child is in a special class, and, if so, what kind; whether the child has ever repeated a grade, and, if so, the grade and reason; and to describe any academic or other problems the child has had in school. Negative answers to each of these items are scored 1, and answers indicating school difficulties are scored 0. These items thus provide three 0-1 scores, which, when added to the score for academic performance, yield a 0-6 score for the School scale.

Construction of the Child Behavior Profile

The behavior problem scales of the profile were derived through factor analysis of CBCLs filled out by parents of 450 boys being evaluated in 20 East Coast mental health settings. These included guidance clinics, health maintenance organizations, and private practices. Racial composition was 79.7% white, 18.7% black, and 1.6% other. Mean socioeconomic status (SES) was 4.4 ($SD = 1.8$) based on Hollingshead's (Note 1) 7-step scale for breadwinner's occupation. To ensure that younger and older boys contributed equally to the profile, the sample contained equal numbers of 6- to 8-year-olds and 9- to 11-year-olds, with approximately equal numbers at each year.

Norms for the profile were computed from CBCLs of 300 normal boys, 50 at each age from 6 to 11. These CBCLs were obtained by interviewers who went to randomly selected homes in the greater Washington, D.C., area, as described elsewhere (Achenbach & Edelbrock, Note 2). The normative sample contained no boys who had received mental health services in the previous year. Racial composition was 79.4% white, 18.3% black, and 2.3% other. Mean SES was 4.1 ($SD = 1.7$).

Results

Behavior Problem Scales

The frequency with which parents endorsed each item (i.e., scored it 1 or 2) was first tabulated to identify items that were too rare or common to add to the discriminative power of factor-based scales. With a lower cutoff of 5% and an upper cutoff of 95%, four items were found to be too rare and none was too common for inclusion in the factor analysis. The four low frequency items were: Item 75. Sexual problems (describe) ____; Item 78. Smears or plays with bowel movements; Item 105. Uses alcohol or drugs (describe) ____; and Item 110. Wishes to be of opposite sex. The low frequency for Items 73 and 110 does not mean that no sexual items remained for analysis (e.g., Item 5. Behaves like opposite sex; Item 59. Plays with own sex parts in public; Item 60. Plays with own sex parts too much; and Item 96. Thinks about sex too much) were all reported for at least 5% of the cases).

Narrow-band scales. A principal components analysis was performed on the 114 items meeting the 5% criterion for the 450 subjects. Because there is no unique criterion for rotation to simple structure, orthogonal (varimax) and oblique (direct quartimin) rotations were both performed on varying numbers of factors to identify the most robust. When more than 13 factors were rotated, factors that had consistently appeared in smaller rotations began to break down into factors having only two or three large loadings. When less than 11 factors were rotated, substantial groupings that occurred in the 11-, 12-, and 13-factor rotations were combined into very large factors. The 12-factor varimax rotation was selected as containing the best representation of the factors that appeared most consistently in the various rotations. However, only the largest 9 of the 12 factors were retained for the profile, as the smallest 3 had only 3-5 items with loadings $\geq .30$. Each of the 9 factors had at least 8 items with loadings $\geq .30$. Because the largest rotated factor (*Aggressive*) had 33 items with loadings $\geq .30$, and many of the items with loadings between .30 and .40 also had substantial loadings on other factors, only the items with loadings $\geq .40$ on this factor

Table 1

First-Order Varimax Loadings on Behavior Problem Scales

Internalizing scales		IV. Obsessive-Compulsive (cont.)	VIII. Aggressive (cont.)
<i>I. Schizoid</i>		93. Excess talk .34	68. Screams
40. Auditory hallucination .55		47. Nightmares .33	90. Swears
70. Visual hallucination .50		50. Anxious .33	25. Poor peer relations
29. Fears .44		Eigenvalue 4.03	88. Sulks
30. Fears school .41		<i>V. Somatic Complaints</i>	7. Brags
11. Clings to adults .37		56f. Stomach problems .64	43. Lies, cheats
50. Anxious .36		56a. Pains .50	27. Jealous
47. Nightmares .31		56b. Headaches .58	87. Moody
59. Public masturbation .30		56c. Nausea .56	19. Demands attention
75. Shy, timid .30		56g. Vomits .44	93. Excess talk
Eigenvalue 2.53		49. Constipated .41	48. Unliked
<i>II. Depressed</i>		51. Dizziness .39	Eigenvalue
35. Feels worthless .68		77. Sleeps much .32	<i>IX. Delinquent</i>
52. Feels guilty .67		54. Overtired .31	82. Steals outside home
32. Needs to be perfect .58		Eigenvalue 3.08	81. Steals at home
33. Feels unloved .55		<i>Mixed scale</i>	21. Destroys things belong-
112. Worrying .52		<i>VI. Social Withdrawal</i>	ing to others
103. Sad .51		48. Unliked .59	106. Vandalism
31. Fears own impulses .48		25. Poor peer relations .59	72. Sets fires
91. Suicidal talk .46		111. Withdrawn .56	101. Truant
12. Lonely .40		42. Likes to be alone .46	67. Runs away
14. Cries much .39		38. Is teased .36	39. Bad friends
50. Anxious .39		64. Prefers younger children .33	43. Lies, cheats
71. Self-conscious .39		34. Feels persecuted .32	20. Destroys own things
34. Feels persecuted .34		102. Slow moving .31	90. Swears
88. Sulks .32		Eigenvalue 3.05	23. Disobedient at school
45. Nervous .31		<i>Externalizing scales</i>	Eigenvalue
89. Suspicious .30		<i>VII. Hyperactive</i>	<i>Other Problems</i>
18. Harms self .30		8. Can't concentrate .65	2. Allergy
Eigenvalue 4.94		1. Acts too young .58	4. Asthma
<i>III. Uncommunicative</i>		61. Poor school work .56	5. Acts like opposite sex
65. Won't talk .61		62. Clumsy .48	6. Encopresis
69. Secretive .50		13. Confused .45	15. Cruel to animals
75. Shy, timid .42		17. Daydreams .43	24. Doesn't eat well
103. Sad .36		41. Impulsive .40	26. Lacks guilt
80. Stares blankly .33		64. Prefers younger children .40	28. Eats nonfood
71. Self-conscious .32		10. Hyperactive .36	36. Accident prone
13. Confused .32		79. Speech problem .31	44. Bites nails
86. Stubborn .30		20. Destroys own things .30	53. Overeats
Eigenvalue 2.97		Eigenvalue 3.75	55. Overweight
<i>IV. Obsessive-Compulsive</i>		<i>VIII. Aggressive</i>	56d. Eye problems
85. Strange ideas .52		3. Argues .71	56e. Rashes
100. Can't sleep .52		22. Disobedient at home .66	58. Picking
76. Sleeps little .45		95. Temper tantrums .64	60. Excess masturbation
84. Strange behavior .43		86. Stubborn .63	63. Prefers older children
9. Obsessions .42		37. Fighting .61	73. Sex problems
92. Walks, talks in sleep .40		16. Cruel to others .60	78. Smears feces
80. Stares blankly .40		97. Threatens people .57	96. Sex preoccupation
17. Daydreams .38		94. Teases .56	98. Thumb sucking
46. Twitches .37		74. Shows off .55	99. Too neat
83. Hoarding .37		104. Loud .51	105. Alcohol, drugs
66. Compulsions .36		23. Disobedient at school .51	107. Wets self
54. Overtired .36		57. Attacks people .50	108. Wets bed
13. Confused .35			109. Whining
			110. Wishes to be opposite sex

Note. Items are designated with the numbers they bear on the Child Behavior Checklist (CBCL) and summary labels for their content. For actual wording of items, see the CBCL. Other Problems items were excluded from scales because of low frequency or low loadings.

were retained for the scale constructed from the factor. Each of the scales constructed from the other 8 factors consisted of the items having loadings $\geq .30$ on those factors. The items, their varimax loadings, eigenvalues, and descriptive labels for each scale are presented in Table 1.

Scoring of scales. Because equal weights for items are likely to provide greater robustness in a linear discrimination system than are weights based on factor loadings (Wainer, 1976), the unweighted raw scores (0, 1, or 2) for all items of a scale were summed to obtain the subject's total score on the scale. The raw scores obtained by the 300 normal boys were used to compute normalized T scores for each of the nine scales. A T score of 80 was assigned to the highest raw score in the normal sample for each scale, excluding outliers. Because a substantial number of clinical subjects obtained higher raw scores than any normal subject on each scale, T scores up to 90 were added to the T scores based on the normal sample. This was done by dividing the T scores from 80 to 90 into as many intervals as there were raw scores between the highest score obtained by a normal subject and the highest raw score obtainable on the scale. These fractional T scores were then assigned to the raw scores from the highest score in the normal sample to the highest possible score and were rounded to the nearest whole T score.

Second-order analysis. The normalized T scores obtained by the clinical sample on the nine scales were intercorrelated and subjected to a principal components analysis, with varimax and direct quartimin rotations of the two factors having eigenvalues > 1.00 . The results of the two rotations were very similar in that Scales 1-5 (Table 1) all had loadings $\geq .63$ on the first second-order factor, whereas Scales 7-9 had loadings $\geq .65$ on the other second-order factor. There was a small difference in the ordering of the scales in that Scale 4 had a slightly higher loading than Scale 5 in the varimax rotation, whereas this order was reversed in the quartimin rotation. Scale 6 had moderate loadings on both factors in both rotations; they were .38 and .44 on the two factors of the quartimin rotation and .44 and .51 on the two factors of the varimax rotation. Because the scoring of some items on

more than one scale might inflate the correlations among scales, the second-order analysis was repeated on first-order scales from which all redundantly scored items were deleted. This yielded the same two second-order factors.

The ordering of the scales in Table 1 follows the order of their loadings on the second-order varimax factors, with Scale I having the highest loading on the first second-order factor, Scale VI having moderate loadings on both second-order factors, and Scales VII-IX having progressively higher loadings on the second second-order factor. The items of Scales I-V clearly form a broad-band grouping like the Personality Problem and Internalizing groupings found previously, whereas the items on Scales VII-IX form a broad-band grouping like the Conduct Problem and Externalizing groupings found previously (Achenbach, 1966; Miller, 1967; Quay & Peterson, 1967).

To obtain normalized T scores for Internalizing and Externalizing, raw scores were computed for each normal subject by summing his scores on all items of the five Internalizing scales and all items of the three Externalizing scales. Items that were included on more than one scale were scored only once to obtain the Internalizing or Externalizing score, but the three items that appeared on at least one Internalizing and one Externalizing scale were counted once each toward both the Internalizing and Externalizing scores. Normalized T scores were derived from the distributions of raw scores for Internalizing and Externalizing in the same way as for the nine first-order scales, with the following modification: Because the range of possible Internalizing and Externalizing scores extended far above the highest score obtained by any of the 450 clinical subjects, the highest raw score actually obtained by any clinical subject was assigned a T score of 89 and all higher possible raw scores were assigned a T score of 90. The T scores from 80 to 89 were assigned by dividing these T scores into as many intervals as there were raw scores from the highest normal (excluding outliers) to the highest clinical subject. These fractional T scores were assigned in sequence to the raw scores ranging from the highest score obtained by a normal subject to the highest score obtained by a

clinical subject. The fractional T scores were then rounded to the nearest whole T score.

Social Competence Scales

The scores obtained on the social competence scales by the 300 normal subjects were used to obtain normalized T scores. A T score of 20 was assigned to the lowest raw score obtained by a normal subject, excluding outliers. Because some clinical subjects obtained lower scores than any normal subjects on these scales, T scores between 10 and 20 were assigned by dividing them into as many intervals as there were raw scores between the score assigned a T score of 20 and the lowest raw score obtainable on the scale.

Age, SES, and Clinical Versus Normal Comparisons

To assess differences in scores related to age, SES, and clinical status, unweighted-means analyses of variance (ANOVAs) were performed on the 300 normal subjects and 300 of the clinical subjects, 50 at each age. SES was divided into three levels, comprising Hollingshead Occupational Categories 1 and 2, 3 and 4 and 5-7. Age was divided into two levels, years 6-8 and 9-11.

Behavior problem scales. A 3 (SES) \times 2 (age) \times 2 (clinical vs. normal) \times 9 (repeated measures on behavior problem scales) ANOVA showed significantly higher scores for clinical than normal subjects, $F(1, 588) = 517.73$, $p < .001$. SES was also significant, with upper-SES subjects obtaining the lowest scores and lower-SES subjects obtaining the highest, $F(2, 588) = 3.37$, $p < .05$. Age showed no effect ($F = .09$), and there were no significant interactions among SES, age, and clinical status. The repeated measures effect of scale was significant, $F(8, 4704) = 8.64$, $p < .001$, as were the interactions of scale with age, $F(8, 4704) = 2.39$, $p < .05$, with clinical status, $F(8, 4704) = 32.61$, $p < .001$, and with age and clinical status together, $F(8, 4704) = 2.33$, $p < .05$.

To elucidate these effects, a 3 (SES) \times 2 (age) \times 2 (clinical status) ANOVA was performed on each scale, with consideration being given only to differences that were reflected in

significant F values for the overall repeated measures ANOVA. All nine of the smaller ANOVAs showed higher scores for clinical than normal subjects, with $F(1, 588)$ values ranging from 67.97 to 408.29, all $ps < .001$. SES was significant for Somatic Complaints, $F(2, 588) = 3.88$, $p < .05$; Hyperactive, $F(2, 588) = 4.44$, $p < .05$; Aggressive, $F(2, 588) = 3.45$, $p < .05$; and Delinquent, $F(2, 588) = 5.43$, $p < .01$. In all significant comparisons, lower-SES subjects had the highest scores and upper-SES the lowest. Modified least significant difference contrasts (Winer, 1971) showed significantly higher scores for lower- than upper-SES subjects on all four scales and significantly higher scores for lower- than middle-SES subjects on the Delinquency scale.

The ANOVA for the Schizoid scale showed significantly higher scores for younger than older boys, $F(1, 588) = 6.33$, $p < .05$, but the lack of significant main effects for age in the other eight ANOVAs and the very small F of .09 for age in the repeated measures ANOVA indicates that age differences were minimal. Significant interactions between age and clinical status in the ANOVAs for the Depressed and Social Withdrawal scales both reflected higher scores for older clinical subjects than younger clinical subjects and lower scores for older normals than younger normals. However, on both scales, clinical subjects of both age groups scored significantly higher than normals. Table 2 presents all the mean scores, collapsed over age to save space.

Internalizing, Externalizing, and total score. Separate 3 (SES) \times 2 (age) \times 2 (clinical status) ANOVAs on Internalizing, Externalizing, and total raw score all showed significantly higher scores for clinical than normal subjects, with $F(1, 588)$ values ranging from 362.47 to 479.76, $ps < .001$. The SES effect was also significant for Externalizing, $F(2, 588) = 3.57$, $p < .05$, and for total score, $F(2, 588) = 4.63$, $p < .01$. In both cases, contrasts showed significantly higher scores for lower-SES than upper-SES boys. No other effects were significant in any of the ANOVAs.

Social competence scales. In a 3 (SES) \times 2 (age) \times 2 (clinical status) \times 3 (repeated measures on social competence scales) ANOVA, clinical subjects had significantly lower scores than normals, $F(1, 528) = 254.78$, $p < .001$.

Table 2
Mean Scores for Clinical and Normal Boys on Scales of the Child Behavior Profile

Scale	Upper SES		Middle SES		Lower SES	
	Clinical (63)	Normal (67)	Clinical (141)	Normal (97)	Clinical (96)	Normal (136)
Social Competence						
Activities	48.2	53.4	45.6	52.9	43.3	49.1
Social	42.1	55.4	39.1	51.9	40.0	50.0
School	43.0	60.1	41.2	57.8	38.2	52.3
Behavior Problems						
Schizoid	62.4	55.7	64.3	53.5	63.3	54.0
Depressed	63.8	52.9	66.1	52.4	67.4	52.5
Uncommunicative	65.0	52.6	67.0	53.1	68.2	53.5
Obsessive-Compulsive	62.5	52.3	64.9	52.4	67.0	53.2
Somatic Complaints	61.0	56.4	62.3	55.9	61.4	57.0
Social Withdrawal	64.5	54.2	65.7	53.7	66.9	53.7
Hyperactive	66.0	50.5	67.4	52.8	68.7	53.5
Aggressive	64.3	51.9	68.2	50.6	70.6	51.5
Delinquent	66.1	54.9	68.1	54.5	69.4	56.6
Internalizing	64.6	51.1	67.2	50.7	68.8	51.4
Externalizing	66.7	50.8	69.3	50.4	71.4	51.6
Total raw score	54.0	20.4	61.8	20.9	65.6	22.8

Note. Numbers in parentheses are *ns*. SES = socioeconomic status. All scores except total raw score are normalized *T* scores. Differences between clinical and normal samples are all significant at $p < .001$. See text for tests of SES differences.

(The total number of cases was less than in the behavior problem ANOVAs because some of the youngest subjects had insufficient school data to be scored on the school scale.) SES was also significant, with lower-SES subjects having the lowest scores and upper-SES subjects the highest scores, $F(2, 528) = 14.61$, $p < .001$. Age differences were not significant ($F = 1.68$), nor were any of the interactions among non-repeated measures dimensions. The repeated measures dimension was significant, $F(2, 1056) = 8.65$, $p < .001$, as were the interactions of scale with age, $F(2, 1056) = 3.09$, $p < .05$, and with clinical status, $F(2, 1056) = 27.83$, $p < .001$.

SES \times Age \times Clinical Status ANOVAs on each of the three social competence scales showed significantly lower scores for clinical subjects than normal subjects on all three scales, with $F_s(1, 528)$ ranging from 42.36 to 179.47, all $p_s < .001$. SES differences were also significant on all three, with $F_s(2, 528)$ ranging from 4.74 to 9.71, all $p_s < .01$. Contrasts showed significantly higher scores for upper-SES than lower-SES subjects on all three

scales and higher scores for upper-SES than middle-SES subjects on the Social scale. Middle-SES subjects scored significantly higher than lower-SES subjects on the Activities and School scales. The only other significant effect reflected higher scores by older than younger subjects on the Social scale $F(1, 528) = 4.73$, $p < .05$.

Format of the Child Behavior Profile

On the computer-scored version of the profile, face sheets describe the nature and purpose of the profile and provide a listing of items on each behavior problem scale, plus items not appearing on any scale. The printout for the behavior problem scales presents a graphic display in which raw scores for the scales are listed in nine columns, percentiles are listed to the left, and *T* scores are listed to the right. An asterisk designates the child's raw score in each column of the display, and the asterisks can be connected by pencil to provide a visual profile. Below the graphic display are printed abbreviations of the items reported by the

parent on each scale, together with the score (1 or 2) given each item by the parent, and the child's raw score and *T* score for each scale. To the right of the nine scales are printed any items scored as present but not belonging to a scale. Also printed are the total number of items scored as present and the sum of 1s and 2s for all items, for the Internalizing items, and for the Externalizing items, plus the *T* scores for Internalizing and Externalizing. The social competence scales are presented in similar fashion on another page. The hand-scored version of the profile is like the computerized version, except that all items are printed and the scorer enters and sums the scores obtained.

Test-Retest Reliability

An interviewer obtained CBCLs from mothers of 12 normal boys on two occasions at a mean interval of 8 days (range = 7-12 days). Pearson correlation coefficients on the 12 scales, Internalizing, Externalizing, and total behavior problem scores ranged from .72 for the Activities scale to .97 for total score (all $ps < .01$), with a mean of .89. (All means of correlations were computed by *z* transformation.) Dependent *t* tests for which *p* values were multiplied by 15 to correct for the number of comparisons (Winer, 1971) showed one significant difference from Time 1 to Time 2, a drop in total behavior problem score, $t(11) = 3.81$, $p < .05$. Even though this was the only significant change in means, it should be noted that 14 of the 15 means decreased, whereas only the mean for Delinquent Behavior increased from Time 1 to Time 2 ($p < .01$ for the proportion of decreases/increases by sign test). There thus appears to be a general tendency to report fewer behavior problems and fewer items indicative of competence on the second occasion, although the differences in most scores were slight.

To assess the short-term stability of profile shapes, a *Q* correlation was computed between each boy's Time 1 profile and his Time 2 profile. This was done by standardizing each of the 12 scale scores within the sample and computing a Pearson correlation coefficient between each boy's 12 scale scores at Time 1 and his 12 scores at Time 2. The mean of the 12 correlations on 12 variables was .86, which

is significant at $p < .001$ whether it is treated as a correlation on 12 observations or as 12 subjects \times 12 scores = 144 observations. To determine whether the correlation could be an artifact of the *Q* approach, a baseline correlation was obtained by pairing each boy's Time 1 profile with every other boy's Time 2 profile except his own. The mean of these 66 correlations was $-.04$, which indicates that the mean correlation of .86 between each boy's Time 1 profile and his own Time 2 profile was not an artifact of the *Q* approach. For workers interested in using the nine behavior problem scales alone, it may be useful to know that the mean *Q* correlation between Time 1 and Time 2 9-scale profiles was .84, whereas between random pairs it was .02. As a measure more sensitive to the similarity between profile elevations, the mean of Cattell's (1949) r_p was .78 between Time 1 and Time 2 12-scale profiles and .74 between the 9-scale profiles. For randomly paired Time 1 and Time 2 profiles, the 12- and 9-scale means were .01 and .02, respectively.

Interparent Agreement

Mothers and fathers of 37 clinic boys independently filled out the CBCL. Pearson correlations between scores obtained from mothers' and fathers' CBCLs on the 12 profile scales, Internalizing, Externalizing, and total behavior problem score ranged from .58 for the Activities scale to .87 for the School scale (all $ps < .001$), with a mean of .74. Dependent *t* tests for which *p* values were multiplied by 15 to correct for chance showed a significant interparent difference only on the School scale, where fathers gave higher scores than mothers, $t(36) = 3.38$, $p < .05$. Across all 15 comparisons, the fathers' mean scores were higher on six and the mothers' on nine ($p > .40$ by sign test).

After standardization of scores on each scale within the sample of 37, the mean *Q* correlation for the 37 pairs of 12-scale profiles was .69. For the nine behavior problem scales, the average *Q* correlation was .74. By comparison, the means of the 666 12- and 9-scale correlations for randomly paired mothers and fathers were $-.02$ and $.04$, respectively. The mean r_p for the wife-husband 12-scale profile pairs was .59 and for the 9-scale profile pairs was .69, com-

pared to .03 and .04, respectively, for the random pairs.

Long-term Stability of Behavior Problem Scores

As part of a follow-up study, 46 parents who had filled out the behavior problem portion of the CBCL when applying to child guidance clinics were asked to complete it again at a mean interval of 14.8 months (range = 9-27 months). The families had received a mean of 11.9 clinical interviews (range = 0-50), but all had terminated with the clinics before the follow-up was begun. To avoid overlap between the initial and follow-up data, the parents were asked to report only behavior problems occurring within the 6 months prior to follow-up, rather than within the previous 12 months, as requested on the initial CBCL. Because of this shorter baseline period, possible regression of scores toward the mean, and "hello/good-bye" effects, changes in scores should not necessarily be interpreted as indicating improvement.

Pearson correlations for the nine scales, Internalizing, Externalizing, and total score ranged from .26 for Somatic Complaints to .79 for Delinquent Behavior, with a mean of .63, all significant except Somatic Complaints. However, the follow-up CBCLs showed decreases on all nine behavior problem scales, as well as on total behavior problem score and *T* scores for Externalizing and Internalizing. With *p* values multiplied by 12 to correct for chance, the decreases were significant by dependent *t* tests for all scores except Schizoid, Somatic Complaints, and Social Withdrawal. The general decrease in reported problems was also reflected in the mean r_p of .43 between intake and follow-up profiles, compared with a mean r_p of .03 between the 1,035 randomly paired intake and follow-up profiles. Despite the decreases in reported problems, the mean *Q* correlation between initial and follow-up *T* scores was .64, indicating considerable long-term stability in profile shape, as compared to the mean *Q* correlation of .02 for the 1,035 random pairs.

Discussion

Although no other studies have taken precisely the same approach, comparison of

the present findings with the most similar previous studies, those of Achenbach (1966) and Miller (1967), reveals considerable similarity along with some differences that are worth noting. (A more extensive survey of previous findings for both sexes, various age groups, and various sources of data is presented by Achenbach & Edelbrock, in press.) Despite the differences in behavior checklists and the fact that Achenbach (1966) used case history data, six of the present narrow-band factors are similar to factors that he obtained for boys. These six are the Schizoid, Obsessive-Compulsive, Somatic Complaints, Hyperactive, Aggressive, and Delinquent factors. The last four of these six are also similar to narrow-band factors found by Miller (1967), which he named Anxiety, Hyperactivity, Infantile Aggression, and Antisocial. In addition, the Social Withdrawal factor found in the present study is similar to Miller's factor of the same name. Miller's failure to find factors like the Schizoid and Obsessive-Compulsive factors is probably due, as Miller pointed out, to the lack of severely disturbed children in his sample.

The remaining two narrow-band factors in the present study, those labeled Depressed and Uncommunicative, have no direct counterparts in the previous factor analyses of boys' behavior problems. However, the Depressed factor is quite similar to the Depressive Symptoms factor that Achenbach (1966) obtained for girls. The emergence of such a factor for boys suggests that cultural changes may be leading either to a greater incidence of depression in boys or to a greater willingness to acknowledge such feelings in boys. Whatever the reason, it appears that the recent spurt of interest in childhood depression is well justified (e.g., Lewis & Lewis, Note 3). The other narrow-band factor, labeled *Uncommunicative*, has no clear counterpart in either of the previous studies. Although the value of each scale lies less in the interpretation of its meaning than in its ability to add discriminative power to the profile as a whole, this scale is suggestive of a constriction in self-expression that might accompany depression in some children.

The second-order Externalizing and Internalizing factors are quite similar to the group-

ings given these names by Achenbach (1966) and to the second-order Aggression and Social Inhibition factors found by Miller (1967). Miller's remaining second-order factor, entitled *Learning Disabilities*, was not likely to appear in the present data because of the different approach taken to the scoring of school performance. Miller's inclusion of several similar items reflecting poor school performance (e.g., reads poorly; spells poorly; writes poorly) made a factor comprising these items almost inevitable. To avoid factors resulting from redundancy in items, the CBCL includes only the general item, *poor school work*, as a behavior problem. However, to provide a differentiated picture of school performance, the CBCL assigns scores for all academic subjects, which are then averaged and combined with scores for special versus regular class status, repeating grades, and other school problems to yield a score for the School scale.

The major objective of this research does not, of course, end with the creation of scales for behavioral problems and competencies. More important is the value of the profile for describing children's behavior in an economical but comprehensive and meaningful fashion, the power of the profile to discriminate among children who may benefit from different kinds of help, and the sensitivity of the profile to changes as well as stabilities in children's behavior. Because they preserve a maximum of information about children's behavior, the profile patterns may provide a much better basis for classifying children than do traditional diagnostic categories or scores on individual scales. The short- and long-term test-retest correlations indicate stability in patterning, and the interparent correlations reflect agreement between parents' perceptions of patterning in their children's behavior. Highly significant differences between normal and clinical subjects on all scales also demonstrate discriminative validity. Studies are now under way to determine whether profile patterns can be identified that significantly differ-

entiate children with respect to long-term prognosis and other clinically relevant characteristics.

Reference Notes

1. Hollingshead, A. B. *Two-factor index of social position*. Unpublished manuscript, Yale University, 1957.
2. Achenbach, T. M., & Edelbrock, C. S. *Behavioral problems and competencies reported by parents of normal and disturbed children aged 4 through 16*. Manuscript in preparation, 1977.
3. Lewis, M., & Lewis, D. O. *An overview of some recent research on depression in children*. Paper presented at the meeting on Depression in Childhood, Berkeley, California, June 1976.

References

- Achenbach, T. M. The classification of children's psychiatric symptoms: A factor analytic study. *Psychological Monographs*, 1966, 80(7, Whole No. 615).
- Achenbach, T. M., & Edelbrock, C. S. The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin*, in press.
- American Psychiatric Association. *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, D.C.: Author, 1968.
- Baker, R. P., & Dreger, R. M. The preschool behavioral classification project: An initial report. *Journal of Abnormal Child Psychology*, 1973, 1, 88-120.
- Cattell, R. B. r_p and other coefficients of pattern similarity. *Psychometrika*, 1949, 14, 279-298.
- Doll, E. A. *Vineland Social Maturity Scale*. Circle Pines, Minn.: American Guidance Service, 1965.
- Miller, L. C. Louisville Behavior Checklist for males, 6-12 years of age. *Psychological Reports*, 1967, 21, 885-896.
- Quay, H. C. Patterns of aggression, withdrawal, and immaturity. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood*. New York: Wiley, 1972.
- Quay, H. C., & Peterson, D. R. *Manual for the Behavior Problem Checklist*. Champaign, Ill.: Children's Research Center, University of Illinois, 1967.
- Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 1976, 83, 213-217.
- Winer, B. F. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Received April 28, 1977 ■

Increasing the Interpersonal Problem-Solving Skills of an Alcoholic Population

James C. Intagliata

Division of Community Psychiatry, State University of New York at Buffalo

Sixty-four male alcoholics admitted into the Alcoholism Treatment Program at the Veterans Administration Hospital in Buffalo, New York, were assigned to either the control or treatment group. Control subjects participated in all standard treatment aspects of the program. Treatment subjects, however, participated in an additional 10 sessions of group therapy structured specifically to improve interpersonal problem-solving thinking skills. Comparisons conducted at the point of discharge (generally 6 weeks after admission) demonstrated that treatment subjects had made significantly greater improvement on a measure of problem-solving thinking than had controls. Further, a comparison of subjects' responses in a structured discharge interview demonstrated that treatment subjects were significantly more likely to anticipate and plan ahead for postdischarge problems than were control subjects. Analysis of the data in the study also revealed that the means-ends problem-solving procedure can reliably discriminate individuals within an adult alcoholic population who differ in their levels of social competence and in the quality of their planning for coping with postdischarge problems. Finally, follow-up at the 1-month postdischarge point indicated that the majority of treatment subjects contacted had made practical use of the problem-solving principles that were taught in the group sessions.

It has been suggested that the capacity to problem solve in real-life situations is one criterion for defining positive mental health (Jahoda, 1958). This suggestion has received increasing empirical support from a number of studies examining the relationship between cognitive interpersonal problem-solving skills and psychological adjustment. These studies have used the means-ends problem solving (MEPS) procedure (Platt & Spivack, 1975) to examine interpersonal problem-solving cog-

nition in a variety of groups including preschool age children (Shure, Spivack, & Jaeger, 1971; Shure, Newman, & Silver, Note 1), young adolescents (Platt, Spivack, Altman, Altman, & Peizer, 1974; Spivack & Levine, 1963), and adults (Platt, Scura, & Hannon, 1973; Platt & Spivack, 1972a; Platt, Spivack, & Siegel, 1975). The results have demonstrated that for each of the age categories, problem-solving cognition is an adaptive thinking ability that successfully discriminates between groups that clearly differ in their level of demonstrated adjustment (e.g., impulsive adolescents at a residential school vs. normal high school controls; adult psychiatric patients vs. normal adult controls). While these studies have demonstrated that problem-solving cognition can discriminate between groups that differ grossly in their level of adjustment, there is evidence that this variable can also discriminate among persons in a homogeneous group who differ only in the degree of their social competence or effectiveness (Platt & Spivack, 1972b; Ziegler & Phillips, 1962). Thus, empirical

This article is based on the author's doctoral dissertation, supervised by Murray Levine and submitted to the Department of Psychology at the State University of New York at Buffalo.

The author wishes to express his appreciation to the entire staff of the Alcoholism Treatment Unit 4-D of the Buffalo, New York, Veterans Administration Hospital for their assistance in conducting this study, and to Louis English who made valuable suggestions with respect to the analysis of study data.

Requests for reprints should be sent to James Intagliata, Division of Community Psychiatry, 462 Grider Street, Buffalo, New York 14215.

research has built a case for the assertion that interpersonal problem-solving cognition plays an important role in healthy human functioning.

The evidence that interpersonal problem-solving skill is intimately related to behavioral adjustment has quite naturally stimulated the development of intervention strategies designed to increase the problem-solving effectiveness of groups deficient in this ability. One of the most ambitious and successful strategies used has been Spivack and Shure's (1973) 12-week problem-solving skills training program for primary-grade children. The goal of this program was to teach children a style of problem solving that could guide them to cope more successfully with everyday problems. Results demonstrated that children who received the training, when compared with controls, showed generation of more alternatives to problems, better consequential thought, and a decided shift away from aggressive solutions to problems. Behaviorally, the trained children showed greater concern for the feelings of other children, were better liked by their peers (i.e., were sought out more), showed greater initiative in the classroom, and had greater autonomy (ability to complete activities and overcome obstacles without adult assistance).

Following these results, several attempts were made to modify this training approach so that it could be extended for use with adult populations that were deficient in interpersonal problem-solving skills. For example, Siegel and Spivack (1973) designed a program to train chronic schizophrenics in interpersonal problem-solving processes. Although they only piloted the program with 11 subjects, the authors were encouraged by the patients' responses and concluded that the approach seemed generally feasible. Platt (Note 2) reported on the use of a problem-solving skills training program with incarcerated heroin addicts. Though subjects were not randomly assigned to groups, the authors felt that the findings, which included a significant relationship between graduating from the cognitive problem-solving skills training program and successful parole performance, suggested both the feasibility and potential effectiveness of their intervention approach.

The present article reports the results of an extensive two-part study designed to extend training in cognitive interpersonal problem solving to an adult alcoholic population. The first study explores the interrelationships in this population between scores on the MEP procedure and (a) the Ziegler-Phillips (1962) social competence scores, (b) scaled vocabulary subtest scores from the Wechsler Adult Intelligence Scale (WAIS), and (c) scores on a discharge interview structured to assess the patient's anticipation of and planning for post-discharge problems. The second study assesses the impact of a 10-session cognitive problem-solving skills training program on subjects' problem-solving thinking and behavior.

A major reason for selecting an alcoholic population with which to work was the relatively successful results of the training program that Platt (Note 2) conducted with heroin addicts. His work suggested the possible relevance of such training programs for addict populations in general. It also seemed reasonable that alcoholics might benefit from a program that emphasized learning the habit of generating a variety of constructive coping behaviors in stressful real-life situations.

Study 1

Method

Subjects. The subjects were 63 male patients admitted to the alcoholism treatment unit (4-D) at the Veterans Administration Hospital in Buffalo, New York, during a 14-week period. The treatment program on this unit involves approximately 6 weeks of inpatient care. Only veterans are eligible for admission. The first 32 consecutive admissions were assigned to the control group. The assignment of 32 consecutive admissions to the treatment group, however, did not take place until 3-4 weeks later. This design was used so that treatment subjects would not begin their problem-solving skills training until all control subjects had been discharged from the program. If control subjects had been allowed to be in daily contact with treatment subjects for 6 weeks, their usefulness as a control group would have been diminished substantially. Although 32 subjects were assigned to the treatment group, one patient left the hospital shortly after attending the first treatment session and was dropped from the study sample. Thus, the treatment group was comprised of 31 subjects.

The subjects ranged in age from 23 to 66. Their average age was 45. With respect to educational background, 27% had completed eight grades or less, 59% had attended high school, 36% had completed high

school, 9% had gone to college, and 5% had received a 4-year college degree. Subjects' scaled scores on the WAIS Vocabulary subtest ranged from 5 to 16. The average scaled score was 10. Of the 63 subjects in the sample, 20% were single and had never been married, 30% were single but had been married previously, and 50% were currently married. In the 3 years preceding their admission to the hospital, 51% had held one steady job, 21% had held and lost several jobs, and 10% had not worked at all.

The control and treatment groups were compared with respect to age, educational background, scores on the Vocabulary subtest of the WAIS (which correlates .87 with the Full Scale WAIS score), marital status, and employment history. Results revealed that these two groups did not differ significantly on any of these variables.

Instruments and procedure. Two weeks after their admission, all subjects were administered the MEPS (Platt & Spivack, 1975).

This procedure makes use of story stems portraying situations in which a need is aroused in the protagonist at the beginning of the story and is resolved by him at the end. The respondent is required to complete the story by filling in those events which might have occurred between the arousal and satisfaction of the hero's need. (Platt & Spivack, 1975, p. 16)

The responses were scored for the number of discrete steps enumerated that effectively enabled story protagonists to reach the resolution points of the stories. The first administration of the procedure involved the random selection of 5 stories from the 10-story MEPS protocol. At the 6-week postadmission point, all subjects were administered whichever 5 stories they had not previously been given. Thus, subjects were administered the MEPS procedure both shortly after recovering from alcoholic detoxification and immediately prior to their discharge. Between the two administrations of the MEPS, treatment subjects completed 10 group training sessions in problem-solving thinking skills. They participated in these sessions in addition to all standard components of the 6-week treatment program. Control subjects participated in all aspects of the treatment program with the exception, of course, of the problem-solving training sessions. The nature of the training sessions will be explained in considerably more detail in Study 2.

On the basis of information gathered at the time of admission, each subject was also assigned a Ziegler-Phillips social competency score. Following the procedure outlined by Ziegler and Phillips (1962), the social competence measure is an additive score based on a subject's age, IQ, educational background, employment history, marital history, and occupational level. Subjects received a score of 0, 1, or 2 in each of these six areas. These six scores were totaled for each subject as an index of his social competence.

Finally, each subject received a score based on his responses to a series of standardized probes in a structured discharge interview administered to each patient by his group therapist shortly before discharge. The interview was structured to assess the comprehensive-

ness of subjects' discharge planning with respect to three key life areas: (a) employment, (b) living situation, and (c) free time. Four basic questions were used to explore the discharge planning that each patient had done in each area. These questions assessed (a) the specific plans a subject had made for dealing with this area of his life, (b) any obstacles that the subject anticipated might interfere with the plans he had made, (c) any alternative plans that the subject had considered before deciding on his chosen plan of action, and (d) any evidence that the subject could present of having already taken action to develop or implement his discharge plans. So that subjects' responses to the discharge interview could be scored, all discharge interviews were tape recorded.

Scoring. Scoring procedures for the MEPS testing were developed by Platt and Spivack (1975), whereas procedures for scoring the discharge interview were developed by me and involved only slight modifications of the MEPS scoring system. The scoring for both the MEPS instrument and the discharge interview was carried out by two senior psychology students who were unaware of the group membership of any individual protocol.

The scoring procedure for the MEPS testing involved noting (a) the relevancy, the number and kinds of means generated by a subject, (b) the number of elaborations or added details that a subject provided to explain each means, (c) the number of obstacles mentioned that might get in the way of the hero reaching the goal in the story, and (d) any mention of time elapsing before the hero reaches the resolution point of the story. A relevant means was defined as an instrumental act that enabled the hero to move toward or reach the goal in the story.

The scoring for the discharge interviews involved noting (a) the number of discrete, relevant means (instrumental acts) that a subject described as his plan for dealing with problems at work, home, or occupying his free time; (b) the number of obstacles mentioned that the subject felt might get in the way of his carrying out his plans, (c) the number of alternative plans that a subject had considered but decided not to put into action, and (d) the number of discrete acts in which a subject had engaged to formulate or implement his discharge plans.

The two raters were trained in the MEPS scoring procedures over a period of 1 month. As a first step, they familiarized themselves with the description of the scoring procedures provided by Platt and Spivack (1975). The final step, following several weeks of practice with the system, was the assignment of 15 hypothetical stories to each of the raters. The reliability coefficient computed between the total number of means assigned by the two raters to each story was .88 (Spearman-Brown prediction formula). Since the scoring procedures for the discharge interview were fundamentally the same as those for the MEPS scoring, no separate reliability coefficient was calculated for the interview scoring. The clear-cut scoring procedures for the Ziegler-Phillips competency measure have already been described in the preceding section and are outlined in greater detail by Ziegler and Phillips (1962).

Results

This study was conducted to explore the content and concurrent validity of the MEPS measure by examining the set of interrelationships between it and measures of subjects' IQ, social competence, and discharge planning. For the purpose of examining these relationships, the following specific scores were used: (a) The total number of relevant means served as a subject's MEPS score; (b) the subject's scaled WAIS Vocabulary score served as an index of IQ; (c) the subject's total score from the six life areas designated by Ziegler and Phillips (1962) served as the social competence measure; and (d) the subject's total points from the areas of work, living situation, and free time planning served as the discharge interview score (1 point for each discrete plan, obstacle, alternative, and action step mentioned).

When the interrelationships between scores on the first administration of the MEPS, the scaled Vocabulary score from the WAIS, and the Ziegler-Phillips social competency measure were explored, it was demonstrated that scores on the MEPS instrument were significantly related both to the social competency score ($r = .20$, $n = 58$, $p < .05$) and to the scaled Vocabulary subtest score ($r = .38$, $n = 62$, $p < .001$). However, when the effect of the Vocabulary score was partialled out of the relationship between the MEPS score and social competency score, the strength of the relationship decreased greatly and was no longer significant ($r = .06$). This result is unlike that reported by Platt and Spivack (1972a). They found that when general intelligence factors were removed from the social competency score, the relationship between the MEPS measure of interpersonal problem-solving cognition and the social competency score increased.

A second set of interrelationships that was explored examined the relationship between the scores on the final administration of the MEPS and both the discharge interview score and the scaled Vocabulary subtest score from the WAIS. Results revealed that the final MEPS score was related quite significantly both to the discharge interview score ($r = .38$, $n = 56$, $p < .002$) and to the original vocabulary measure ($r = .26$, $n = 55$, $p < .02$). When

the effect of the vocabulary score was partialled out of the relationship between the final MEPS score and the discharge interview score, the relationship between the two remained strong and significant ($r = .33$).

When the relationship between the WAIS vocabulary measure and the MEPS scores on both the initial and final MEPS testing was examined for the entire sample, there was a significant correlation between WAIS and MEPS scores at both points ($r = .38$ for initial, $r = .26$ for final). However, when relationships were examined separately for the control and treatment groups, it appeared that the problem-solving training sessions, to which only the treatment subjects were exposed, may have had some impact on the way in which the MEPS and vocabulary measures relate. For the control group, the correlation between the two measures changed from .43 (initial MEPS and vocabulary) to .32 (final MEPS and vocabulary). Both of these correlations were significant ($p < .007$ and $p < .04$, respectively). For the treatment group, on the other hand, the correlation coefficient changed from .36 (initial MEPS and vocabulary) to $-.05$ (final MEPS and vocabulary). Although the relationship between the two measures was significant for the treatment group at the initial MEPS testing ($p < .02$), the two measures were relatively unrelated at the time of the final MEPS administration ($p < .30$). Further, not only did the strength of the relationship decrease for the treatment group, but the valence shifted from positive to negative.

Study 2

Method

Subjects. The subjects are the same as described in Study 1.

Instruments and procedure. The treatment group of 31 subjects participated in a series of 10 60-minute sessions of interpersonal problem-solving group therapy. These sessions comprised a systematic intervention intended to increase participants' problem-solving skills. The 10 sessions combined and integrated a variety of materials prepared by Platt, Spivack, and Swift (1975) in a structured program of interpersonal problem-solving group therapy for adults. The 10 sessions were organized to teach a four-step approach to interpersonal problem solving. These steps were to (a) recognize that a problem exists; (b) define the problem; (c) generate a number of alternative solutions; and (d) select the best alternative after having looked ahead to imagine

the likely consequences of each. The treatment group of 31 subjects was divided into four subgroups (7, 6, 8, 10), each of which began the 10-session program approximately 1 week apart. The 10 sessions took place over a period of 4 weeks (2-3 sessions per week). All participants had spent at least 2 weeks in the treatment program before problem-solving sessions began.

As explained in Study 1, all subjects were administered the MEPS procedure 2 weeks following their admission to the hospital and again 4 weeks later. This arrangement provided a means of assessing the impact of the problem-solving sessions on the participants' problem-solving thinking skills. The timing was set so that all members of the treatment group had completed the 10-session program prior to the final MEPS administration.

In addition to the problem-solving sessions, one other independent variable was manipulated in this study. This variable was the preparation set of subjects who were about to be administered the structured interview to assess the quality and comprehensiveness of their discharge planning. One half of the subjects in both the control and treatment groups were randomly assigned to the condition in which they were told that their group therapist would interview them about the planning they had done for their discharge. They were so informed about an hour before they were to be interviewed. The other half of the subjects in each group was assigned to the condition in which they were not informed of the topic of the interview until it actually began. All of the subjects who were informed about the topic of their approaching interview were told by me that the interview would focus on their discharge

planning and would give them a chance to demonstrate to their group leader that they had given a lot of thought to the problems that they would soon be facing.

This manipulation was designed to serve two purposes. First, it provided a means of testing the efficiency of the problem-solving skills training as a means of increasing the effectiveness and thoroughness of subjects' planning for the real-life problems posed by their discharge. If the mere suggestion to control subjects that they ought to give a little thought to the problems that they would face following discharge should result in their performing as well as treatment subjects who have completed the problem-solving sessions, one would have to conclude that the sessions were a rather inefficient means of encouraging subjects to anticipate problems and plan ahead. Second, this manipulation provided a means of testing the degree to which the generalization of problem-solving skills taught in the training sessions was dependent on specific reminder cues.

Thus, the second part of the study provided two measures to assess the impact of the problem-solving training sessions. The first was the paper-and-pencil MEPS measure used to assess changes in cognitive problem-solving behavior. The second was the structured discharge interview to assess the quality of subjects' plans for dealing with impending real-life problems.

Results

MEPS measure. It was hypothesized that the treatment group would show significantly

Table 1
Pre-Post Comparisons of Story Elements

Statistic	Pretest			Posttest			Pre-post change		
	Control (32)	Treatment (31)	+ score	Control (27)	Treatment (29)	+ score	Control (27)	Treatment (29)	+ score
No. means							.29	3.50	
<i>M</i>	5.96	5.77		5.85	9.27		3.51	3.83	-3.32**
<i>SD</i>	3.33	2.61	.26	3.62	3.29	-3.70***			
Relevancy score							.06	.14	-1.18
<i>M</i>	.69	.76	-1.18	.75	.91	-2.70*	.27	.25	
<i>SD</i>	.24	.19		.26	.16				
Enumerations							1.22	2.24	-.80
<i>M</i>	2.25	1.96	.68	3.13	4.24	-.89	4.31	4.73	
<i>SD</i>	2.92	2.41		4.02	4.73				
Obstacles							-.07	.24	.15
<i>M</i>	.21	.16	.63	.07	.41	-1.50	.38	1.02	
<i>SD</i>	.42	.52		.26	1.12				
Time							-.03	.37	.06
<i>M</i>	.18	.03	1.74	.07	.41	-1.60	.43	1.01	
<i>SD</i>	.47	.18		.26	1.02				

Note. Numbers in parentheses are *ns*.

* $p = .01$.

** $p = .002$.

*** $p = .001$.

Table 2
Interview Analysis of Variance

Source	SS	df	MS	F
Main effects	188.4	2	94.2	5.7**
Control vs. treatment (A)	183.7	1	183.7	11.1***
Interview preparation (B)	5.7	1	5.7	.3*
A \times B	.4	1	.4	.02*
Residual	971.5	59	16.4	
Total	1,160.4	62	18.7	

* $p < .99$.

** $p < .006$.

*** $p < .002$.

more improvement on the MEPS measure than would the control group. This was expected, since the MEPS instrument is designed to measure interpersonal problem-solving skills and only treatment subjects participated in group sessions specifically designed to increase these skills. Table 1 presents the means and standard deviations for the various criteria on which the subjects' stories were scored. First, the two groups were compared with respect to the number of relevant means that they generated for the five MEPS stories that were administered as pretests and posttests. An examination of the difference scores (difference = number of posttest means minus number of pretest means) for the two groups shows that the treatment group made significantly greater improvement than did the control group, $t(54) = -3.32$, $p < .002$. Even though the two groups performed approximately the same on the pretest (control = 5.96 means, treatment = 5.77 means), the treatment group generated far more means than did the controls on the posttest (controls = 5.85 means, treatment = 9.27 means).

An examination of the scores obtained by the treatment and control groups on the individual MEPS stories for the posttest demonstrated that the significant overall differences were not accounted for by any one story or by only a few stories. In fact, the average number of means generated by the treatment group was greater than that of the control group for each of the 10 stories that comprise the MEPS procedure.

Mean relevancy scores for the treatment and

control groups are also presented in Table 1. The relevancy score was obtained by dividing the number of relevant means generated by the total of all the means generated for each subject. The relevancy score is an index of the extent to which subjects in the two groups responded with relevant and effective solutions to the MEPS stories administered. Although the treatment group showed more positive change in the relevancy score from pretesting to posttesting than did the control group, this difference did not reach statistical significance, $t(54) = .24$, $p < .24$. That the treatment group increased their relevancy score from .76 to .91, whereas controls increased merely from .69 to .75, however, suggests that there is a strong trend that favors the treatment subjects in the predicted direction.

In addition to the comparisons between the two groups on the basis of number and relevancy of means generated, Table 1 also presents a comparison of the groups with respect to other story elements including enumerations (elaborations) of means, obstacles, and the acknowledgment of passed time. The control and treatment groups did not differ significantly on any of these measures. However, the treatment group tended to perform better on each of these dimensions than did the control group of the posttest protocols.

Discharge interview. Performance in the interview was analyzed by means of a 2×2 analysis of variance (control/treatment vs. prepared for interview/unprepared). The results of this analysis are presented in Table 2. These results demonstrated a highly significant main effect due to group membership, $F(1, 59) = 10.2$, $p < .003$, no main effect resulting from differences in preparation for the interview, $F(1, 59) = .002$, $p < .99$, and no significant interaction between group membership and interview preparation $F(1, 59) = 1.0$, $p < .99$.

Table 3 presents the actual means and standard deviations of interview scores for both the treatment and control groups. This table demonstrates that the significant difference between the two groups in the interview was not a result of one question or a specific subset of questions. Rather, the treatment group performed better on all dimensions of the interview.

Table 3
Comparison of Discharge Interview Performance

Variable and group	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i> <
Total interview score	7.6	3.1			
Control	11.0	4.7	-3.4	51	.002*
Treatment					
Total plans score	4.6	1.7			
Control	5.9	2.2	-2.5	61	.01
Treatment					
Total obstacles score	.8	.8			
Control	1.4	1.2	-2.5	51	.01*
Treatment					
Total plans rejected score	.4	.6			
Control	.9	1.3	-1.7	41	.09
Treatment					
Total action-steps taken score	1.8	1.3			
Control	2.8	1.6	-2.7	61	.008
Treatment					
Total score for employment planning	2.3	1.1			
Control	3.9	1.7	-2.8	51	.007*
Treatment					
Total score for living situation planning	1.7	1.1			
Control	2.7	1.6	-2.8	53	.007*
Treatment					
Total score for free time planning	3.4	1.9			
Control	4.7	2.1	-2.5	61	.01
Treatment					

Note. Numbers in parentheses are degrees of freedom.

* Adjusted for nonhomogeneity of variance.

Though the interview was highly structured, it seemed important to monitor whether the group therapists administered it to both control and treatment subjects in a substantially similar manner. To check this, four of each therapist's discharge interviews were randomly selected for analysis (two interviews of control subjects, two of treatment subjects). Since there were four therapists, this meant that a total of eight interviews were selected from both the control and treatment groups. These interviews of control and treatment groups were compared on each of the following variables for both interviewers and interviewees: (a) number of responses, (b) length of responses (in seconds), and (c) total time taken for the interview. There were no significant differences between interviews of control and treatment subjects for any of the indices examined.

Discussion

The results of Study 1 provide further evidence of the content and concurrent validity

of the MEPS measure. First, MEPS scores were found to be significantly related to the social competency of the alcoholic subjects in the study sample. Further, the MEPS scores were also found to be significantly related to subjects' performance in an interview designed to assess the quality of their planning for dealing with postdischarge problems.

As in previous work reported by Platt and Spivack (1975), the correlation between the MEPS scores and IQ measures reached significance, though it was of a magnitude suggesting that the MEPS is not merely another IQ test. However, unlike the study reported by Platt and Spivack (1972b) in which partialing IQ out of the correlation between MEPS and social competency scores strengthened the relationship, the same partialing procedure in this study significantly weakened the relationship between the two measures. One possible explanation for this discrepancy is that in the Platt and Spivack (1972b) study, the index of IQ was a comprehensive full scale score. In the present study, a scaled vocabulary score

alone was the index of subjects' IQ. Thus, the subjects' verbal skills, and not their general intellectual ability, was the factor that was partialled out. Since the MEPS is a measure of cognitive ability heavily dependent on a subject's ability to verbalize problem-solving thoughts, it is not surprising that the partialing out procedure had a more significant impact in this study.

Finally, Study 1 explored, for the first time, the impact of a problem-solving skills training program on the relationship between the MEPS measure and an index of intellectual ability. The results suggest that the extent to which a subject's score on the MEPS procedure is dependent on verbal skills *decreases* following problem-solving training. Though training did not increase the participants' verbal proficiency, it apparently changed the way in which they used whatever verbal skills they already possessed. Thus, they were able to conceptualize qualitatively better solutions to the real-life problems presented to them in the final MEPS administration than they had prior to training. This is an exciting discovery because it suggests that individuals with even limited verbal skills can learn a more effective cognitive strategy for approaching everyday problems.

Study 2 demonstrated that 10 structured group sessions focusing on the processes of effective interpersonal problem solving were sufficient to result in the treatment group showing significantly more improvement, relative to controls, on the MEPS measure of interpersonal problem-solving thought. A closer look at the specific nature of this greater improvement revealed that the thinking of treatment subjects improved qualitatively as well as quantitatively. On the posttest, treatment subjects generated stories that were longer and contained significantly more effective problem-solving means than those generated by controls. In addition, the proportion of the means generated that were judged to be relevant and effective responses was substantially greater for the treatment group than it was for the control group.

Treatment subjects, however, not only performed better on a paper-and-pencil measure of problem-solving thought, but they also utilized their improved problem-solving think-

ing to produce qualitatively better plans for dealing with postdischarge problems than did controls. In the structured discharge interview, treatment subjects demonstrated that the positive gains they had made following problem-solving training generalized successfully at least to an in-hospital situation in which they were asked to confront real-life problems.

The population of alcoholic subjects in this study demonstrated a significant deficiency in interpersonal problem-solving skills when compared to the MEPS norms for a normal adult population provided by Platt and Spivack (1975). As discussed previously, Platt (Note 2) identified a population of heroin addicts as similarly deficient (relative to normals) in these same cognitive skills. Together these findings suggest that drug-addict populations may, in general, be very appropriate targets for problem-solving skills training. More importantly, the generally positive results of such training reported by Platt (Note 2) and in the present studies provide encouragement that such problem-solving skill deficiencies may be amenable to positive change. Future research in this area, however, must do more than explore the feasibility and potential effectiveness of problem-solving training with addict populations. Studies must attempt to demonstrate that this intervention approach can lead to significant positive behavioral change after subjects have returned to the community. Designs that include plans for moderate and long-term follow-up of subjects are essential.

The design of the present study included only a modest follow-up component. Treatment subjects were contacted 1 month following their discharge to determine how well they remembered the problem-solving principles that they were taught and whether or not they had made use of these principles in dealing with real-life problems. Although the majority of those contacted (14/22) reported having made practical use of the principles, it was evident that subjects had already forgotten significant portions of the training material. Thus, if more rigorous follow-up efforts are to be conducted in the future, it would seem reasonable to also explore the utility of occasional "refresher" sessions for the subjects receiving problem-solving training. In addi-

tion, to increase the likelihood of long-term impact, it would seem essential to integrate interpersonal problem-solving principles and concepts into all the components of a comprehensive treatment program rather than to artificially restrict problem-solving training to 10 special sessions as was done in the present study.

Certainly the use of problem-solving skills training ought not to be restricted to addict populations. Siegel and Spivack (1973), for example, reported encouraging results with pilot efforts to use such training with chronic schizophrenics. Platt (Note 3) reported that he was aware of some encouraging pilot work using cognitive problem-solving skills training with the mentally retarded. In addition, D'Zurilla and Goldfried (1971) reported that informal clinical experimentation and pilot work with college freshmen suggested that teaching a general cognitive approach to solving real-life problems seemed to be a promising conceptual strategy in aiding these individuals in coping with the transitions to college living.

In selecting populations and designing programs to improve cognitive interpersonal problem-solving skills, however, a number of factors that can affect the outcome of these interventions should be considered. One such factor may be the verbal skills of the participants. It seems reasonable to assert that a certain minimal level of verbal skill is essential for the cognitive training approach to have a significant impact. This assertion is supported, in part, by the results of this study. Although the majority of those contacted in the follow-up reported having made use of the training in real-life problem situations, it was the more verbally skilled subjects who reported having made greatest use of the training principles since leaving the hospital. Thus, although training may have had a significant immediate impact on the treatment group as a whole, the retention and use of the principles taught may be dependent on the level of the subjects' verbal skills. Other important factors that should be considered include the participant's level of social competence; use of the training intervention as an adjunct therapy (as in this study) versus its use as a solo intervention;

and the number, length, and sequencing of training sessions.

Spivack (1973) proposed that healthy psychological functioning is, in large part, dependent on the ability to cope effectively with the demands of problematic life situations. To deal effectively with such demands, an individual must be able to respond in a competent manner both cognitively and behaviorally. That is, a person must be able to decide what is the best way to cope with the problem as well as be able to actually perform the chosen behavior. The treatment intervention used in this study attempted to provide participants with an opportunity to learn and practice both cognitive and overt behavioral problem-solving skills. The focus of the training sessions, however, was clearly to teach participants a cognitive strategy for dealing constructively with all sorts of problems regardless of their specific content. Thus, general cognitive style rather than specific behavioral skills was emphasized. Though this particular emphasis led to desirable consequences for the subjects in this study, other populations may receive more benefits from training programs that focus more on teaching specific effective behaviors for specific types of real-life problems. For example, Argyle, Trower, and Bryant (1974) reported that they have taken the more specific behavioral approach in their work with neurotic psychiatric patients. Further research must be done to determine which syntheses of cognitive and behavioral skills training are most effective for specific problem populations.

Summary

These studies demonstrated three key points: (a) Problem-solving thinking skill is related significantly to social competency and to "planning ahead for problems" behavior in an alcoholic population, (b) problem-solving thinking skills (as measured by MEPS) of adult alcoholics can be improved significantly through the use of structured training sessions, and (c) improved problem-solving thinking is generalized by subjects from the training sessions into real-life problem situations both within the hospital and after their discharge.

Reference Notes

1. Shure, M., Newman, S., & Silver, S. *Problem solving among adjusted, impulsive, and inhibited Head Start children*. Paper presented at the meeting of the Eastern Psychological Association, Atlantic City, New Jersey, April 1970.
2. Platt, J. J. *Evaluation of a correctional treatment program for youthful heroin addicted offenders*. Unpublished manuscript, Hahnemann Medical College and Hospital, Philadelphia, 1975.
3. Platt, J. J. Personal communication, 1976.

References

- Argyle, M., Trower, P., & Bryant, B. Social skills training for neurotic mental patients. *Proceeding of the First International Symposium of Non-Aspects and Techniques of Psychotherapy*, 1974, 1-14.
- D'Zurilla, T. J., & Goldfried, M. R. Problem-solving and behavior modification. *Journal of Abnormal Psychology*, 1971, 78, 107-126.
- Jahoda, M. *Current concepts of positive mental health*. New York: Basic Books, 1958.
- Platt, J. J., Scura, W., & Hannon, J. R. Problem solving thinking of youthful incarcerated heroin addicts. *Journal of Community Psychology*, 1973, 1, 278-291.
- Platt, J. J., & Spivack, G. Problem solving thinking of psychiatric patients. *Journal of Consulting and Clinical Psychology*, 1972, 39, 148-151. (a)
- Platt, J. J., & Spivack, G. Social competence and effective problem-solving thinking in psychiatric patients. *Journal of Clinical Psychology*, 1972, 28, 3-5. (b)
- Platt, J. J., & Spivack, G. *Manual for the means-ends problem solving procedure (MEPS): A measure of interpersonal cognitive problem-solving skill*. Philadelphia, Pa.: Hahnemann Community Mental Health/Mental Retardation Center, Department of Mental Health Sciences, Hahnemann Medical College and Hospital, 1975.
- Shure, M., Spivack, G., & Jaeger, M. Problem-solving thinking and adjustment among disadvantaged pre-school children. *Child Development*, 1971, 42, 1791-1803.
- Siegel, J. M., & Spivack, G. *Problem solving therapy: A new program for chronic schizophrenic patients*. (Research and evaluation report 23) Philadelphia, Pa.: Department of Mental Health Sciences, Hahnemann Medical College and Hospital, 1973.
- Spivack, G. *A conception of healthy human functioning*. (Research and evaluation report.) Philadelphia, Pa.: Department of Mental Health Sciences, Hahnemann Medical College and Hospital, 1973.
- Spivack, G., & Levine, M. *Self-regulation in acting-out and normal adolescents* (Report M-4531). Washington, D.C.: National Institutes of Health, 1963.
- Spivack, G., & Shure, M. *Social adjustment of young children: A cognitive approach to solving real-life problems*. San Francisco: Jossey-Bass, 1973.
- Ziegler, E., & Phillips, L. Social competence and the process-reactive distinction in psychopathology. *Journal of Abnormal and Social Psychology*, 1962, 65, 215-222.

Received May 2, 1977 ■

Preparation for a Stressful Medical Procedure: Effect of Amount of Stimulus Preexposure and Coping Style

Robert H. Shipley
Department of Psychiatry
University of Missouri

James H. Butt
Department of Medicine
University of Missouri Medical Center—
Harry S. Truman Memorial Veterans Hospital

Bruce Horwitz and John E. Farbray
University of Missouri

Anxiety during the stressful medical procedure of endoscopy was studied as a function of the number of prior viewings of an explicit preparation videotape and of repression-sensitization coping style. Sixty naive patients viewed a videotaped endoscopy either zero, one, or three times. Dependent measures included heart rate, behavioral ratings, tranquilizer required, and self-report. On each dependent measure, three viewings generally resulted in the least distress; one, more distress; and zero, the most distress. Most comparisons reached statistical significance. These results are interpreted as resulting from extinction and/or habituation of anxiety. The repression-sensitization factor interacted with heart rate change. Sensitizers showed a monotonic decrease in heart rate as a function of number of tape exposures. Repressors showed an inverted-U-shaped function, with one viewing producing the highest heart rate; this is interpreted as resulting from a disruption of repressing defenses by one tape exposure followed by extinction of fear by three exposures.

Preparation-for-stress messages have generally proved effective in reducing the emotional trauma of stressful real-life experiences (e.g., Cassell, 1965; Janis, 1958; Johnson & Leventhal, 1974; Vernon & Bailey, 1974). The content, emphasis, and medium of the preparation messages have varied greatly, but most are based on the modeling and/or accurate information theories.

Modeling of approach responses has been effective in reducing phobic fear and avoidance

(Bandura & Menlove, 1968). Viewing of models who are initially fearful and who then overcome their fear ("coping model") has been found to be more effective than the viewing of models who are fearless throughout the stressful event (Kazdin, 1974; Meichenbaum, 1971; Vernon, 1974). Melamed and Siegel (1975) found that children who viewed a coping model were subsequently less anxious before and after surgery than children who viewed an unrelated control film.

Preliminary results of this study were reported at the convention of the Association for the Advancement of Behavior Therapy, New York City, December 1976.

The authors wish to acknowledge the help of the professional staff of the Division of Gastroenterology, Department of Medicine, University of Missouri, and the staff of the Medical Media Production Service, Harry S. Truman Memorial Veterans Hospital. Appreciation is also given to Ellen Horwitz and Dennis McGuire for critical readings of the manuscript.

Requests for reprints should be sent to Robert H. Shipley, who is now at the Veterans Administration Hospital, Fulton Street and Erwin Road, Durham, North Carolina 27705.

A second theory of preparation effects focuses on the formation of accurate cognitive expectancies. Emotional distress during a stressful event is thought to be caused by a discrepancy between current experience and prior expectation (Hebb, 1946; Johnson, 1973; McClelland, 1951). Preparation-for-stress messages based on this theory attempt to provide the information necessary for the subject to form accurate expectancies regarding the stressor and their reaction to it (e.g., Johnson & Leventhal, 1974).

The present study used a stressful medical

examination to investigate a third variable: the amount of prior nonreinforced exposure to the stimuli surrounding a stressful event. The histories of most adults include many painful experiences within a medical setting. Stimuli that precede these painful experiences may become conditioned stimuli capable of eliciting anxiety. One of the most reliable formulations coming out of the experimental laboratory is Pavlov's original finding that presentation of the conditioned stimulus in the absence of the unconditioned stimulus leads to extinction of the learned response. Extinction of conditioned fear has been shown to be a positive function of the number and duration of exposures to the conditioned stimulus (Shipley, 1974). Fear can also be generated by exposure to novel stimuli, and it is likely that novel medical procedures elicit this unconditioned fear. As with conditioned fear, the rate of habituation to novel stimuli has generally been found to be a function of the number and duration of stimulus exposures (Graham, 1973).

Extrapolating from this research to the preparation-for-stress area, it was hypothesized that the more exposures a person receives to a preparation message, the less fear will be experienced during the actual stressful event. The amount of stimulus preexposure is not viewed as an alternative to the provision of accurate information or to modeling of coping responses. Instead, it represents another factor that may contribute to the observed beneficial effects of preparation routines and that may be used to design more effective preparation messages.

In the present study, amount of stimulus preexposure to a real-life threatening medical examination was manipulated. Patients scheduled to receive an upper-gastrointestinal endoscopy viewed a videotape of an actual endoscopy either zero, one, or three times; an unrelated control tape was viewed by the "zero" control group. An attempt was made to control the effects of accurate information by providing all subjects with extensive verbal information about the endoscopy. Imitation of coping responses was minimized by use of a model who appeared fearful throughout the endoscopy.

The endoscopy examination seems to be well suited to the study of real-life stress (Johnson

& Leventhal, 1974; Johnson, Morrissey, & Leventhal, 1973). It involves the insertion of a flexible fiberoptic endoscope, 12 mm in diameter, through the mouth and into the gastrointestinal tract. Air is pumped into the gut to make its interior more visible, and the endoscope is manipulated for approximately 15-30 minutes while the physician views the lining of the gastrointestinal tract. Patients judged to be particularly anxious are sedated prior to the examination, but they are not anesthetized, since their active cooperation contributes to the success and safety of the examination. Patients view endoscopy as stressful, and it is easier to study than stressful procedures such as surgery because endoscopy is more time limited. Since patients remain awake during the examination, they are able to demonstrate and assess their distress. The fact that the patient must lie relatively still allows reliable measurement of physiological arousal.

Prior research has suggested that the effect of a preparation message may depend on the individual's characteristic method of coping with stress (Andrew, 1970; DeLong, 1971). The present study included a questionnaire measure of repression-sensitization coping style. Repression-sensitization is a unidimensional categorization, with repressors at one end of the continuum and sensitizers at the other. Sensitizers are generally described as handling stress by being vigilant, overtly anxious, alert to threatening cues, and by using intellectualization as a defense. They actively seek information about a stressor as a means of preparing to experience it. Sensitizers were expected to be initially high in anxiety and to show a monotonic decrease in anxiety as a function of the number of exposures to the preparation videotape. Repressors, on the other hand, are generally described as being overtly nonanxious and as dealing with the threat of impending stress by not thinking about it, repressing it, or denying its potential stressfulness.

It was hypothesized that repressors would be initially low in anxiety but that one exposure to an explicit videotape of the endoscopy examination would reduce their repressing defenses, resulting in increased distress. Repressors were expected to respond to additional exposures to the preparation tape with a

decrease in anxiety similar to that predicted for sensitizers. Thus, repressors who viewed the preparation tape once were expected to demonstrate greater anxiety than those viewing an unrelated control tape. Repressors who viewed the tape three times were expected to show less anxiety than those who received only one exposure to the tape.

Method

Subjects

The subjects were 60 hospitalized patient volunteers who had received no prior endoscopy examinations and who had been scheduled to receive an upper-gastrointestinal endoscopy at the Harry S. Truman Memorial Veterans Hospital in Columbia, Missouri. The 60 subjects consisted of 50 males and 10 females, ranging in age from 22 to 80, with a mean age of 53.0. Twenty subjects were randomly assigned to each experimental condition. Patients were not invited to participate in the study if they could not read or if they were judged to be disoriented or physically unable to complete the experimental procedures.

Measures of Anxiety

Anxiety is considered to be a multidimensional construct that may be reflected in physiological response, observable behavior, and self-report (Lang, 1968). Measures were selected to assess each of these response classes.

Physiological measures. During the endoscopy, heart rate was monitored with a Hewlett Packard Model 7788A polygraph using three fluid column electrodes. Respiration rate and skin conductance were also monitored. Respiration was not analyzed. Skin conductance data are not reported due to questionable validity stemming from inconsistent electrical grounding of the different endoscopes used and to attenuation of electrodermal responding by the atropine administered to all subjects (Lader & Montagu, 1962).

Behavioral measures. A physician-nurse anxiety rating scale, completed after the endoscopy examination by both the patient's endoscopist and nurse, was designed to assess the patient's fear during three time periods: (a) prior to insertion of the endoscope (before scoping), (b) while the endoscope was in the gastrointestinal tract (during scoping), and (c) after removal of the endoscope (after scoping). The scale consists of eight items (e.g., "Patient appears anxious before the tube is passed—gritting teeth, tight muscles, sweating, etc.") to be rated on a 5-point scale ranging from "not at all" to "very much." The item ratings made by the physician and nurse were combined, since Pearson product-moment correlations between the independent ratings were moderate and significant, $r(58) = .44-.55$, $p < .001$. Three summary anxiety ratings were then obtained by averaging the items within the before,

during, and after scoping time periods. The internal consistency of these three subtests ranged from .80 to .87, using Cronbach's (1951) alpha.

A second behavioral indicator of anxiety was whether the patient received diazepam (Valium) intravenously prior to an attempt to pass the endoscope. Only those patients judged by the physician to be highly anxious were given diazepam.

Self-report measures. Two questionnaires were used to obtain self reports of anxiety: the Spielberger State-Trait Anxiety Inventory (STAI) and the Post-Endoscopy Interview Schedule. The STAI (Spielberger, Gorsuch, & Lushene, 1970) consists of two separate 20-item scales. The Trait Anxiety (A-Trait) scale assesses anxiety proneness, and the State Anxiety (A-State) scale measures momentary or situational anxiety. The A-State scale was expected to reflect changes in anxiety level over time and was administered prior to the experimental manipulations as well as before and after the endoscopy. Scores on the A-Trait scale were not expected to differ between groups or to change over time (Spielberger, Auerbach, Wadsworth, Dunn, & Taubee, 1973). It was administered after the endoscopy primarily for descriptive purposes and to allow for calculation of correlations between measures of trait anxiety, state anxiety, and repression-sensitization.

The Post-Endoscopy Interview Schedule was used in a structured inquiry into the patient's reaction to the examination. Consisting of both open-ended questions and items to be rated on a 5-point scale, it was used to obtain a retrospective self-assessment of anxiety during the preparation videotape, before the endoscopy examination, and during the examination. Information was also gathered on the degree of annoyance and physical discomfort experienced during the examination and prior hospitalizations.

Measure of Coping Disposition

A Modified Repression-Sensitization (R-S) Scale, developed by Epstein and Fenz (1967), was used to classify patients as to their characteristic way of dealing with anxiety-provoking stimuli. The Modified R-S scale was developed in an attempt to eliminate the high (around .90) correlation between Byrne's Minnesota Multiphasic Personality Inventory derived R-S scale and measures of anxiety (Byrne, 1964).

Procedure

Each patient in the study, regardless of treatment condition, routinely received information about the endoscopy on three separate occasions—from the physician, the nurse, and the experimenter. The patient was told of the need for the examination, the procedure that would be followed, likely subjective sensations, and that would be followed, likely subjective sensations, and that would be followed, likely subjective sensations. The evening possible but improbable complications. The evening before the endoscopy, the patient was asked to participate in the research study and to sign a consent form. The Spielberger A-State scale and the Modified R-S scale were then completed, and, depending on prior random assignment, the patient was shown the control videotape once, (Group E0), the experimental tape videotape once, (Group E1), or the experimental tape three times

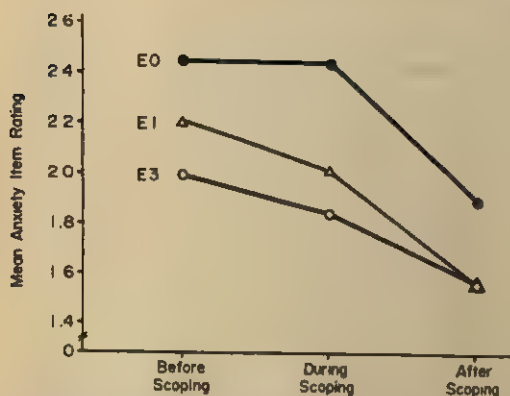


Figure 1. Mean physician-nurse anxiety rating before, during, and after scoping for each of the treatment conditions. (E0, E1, and E3 refer to the groups that viewed the control tape once, the experimental tape once, or the experimental tape three times, respectively.)

(Group E3). Subjects were instructed to watch each tape in its entirety. (Monitoring of time spent looking directly at the tape confirmed that subjects actually received the amount of exposure intended.)

The 18-minute experimental videotape showed a 35-year-old white male patient actually receiving an endoscopy. The patient in the tape showed an "average" amount of distress during the examination, gagging several times and requiring a moderate amount of "calming talk" by the nurse. The 26-minute control videotape was titled *Because I was too Young: The Memoirs of a Distinguished Indiana Physician*.¹

On the morning of the endoscopy, each patient received an intramuscular injection of .6 mg of atropine to reduce oral and stomach secretions, and the A-State scale was completed for the second time. Between 5 and 30 minutes later, the patient was taken into the endoscopy area, and electrodes were attached for the physiological measurements. After the physiological recordings stabilized, a 2-minute baseline was recorded.

Two physicians and a nurse then joined the experimenter in the endoscopy area. All present were blind as to the patient's experimental condition. The patient's throat was sprayed with a local anesthetic (1% Pilocaine), and, if judged necessary by the physician, Valium was administered. For those subjects who received Valium, the dosage ranged from 1.0 mg to 15.0 mg, with a mean of 7.7 mg. The physiological data were recorded for at least 1 minute prior to first placing the endoscope inside the patient's mouth and for 10 minutes thereafter. The experimenter also recorded (a) number of attempts necessary to insert the endoscope past the superior esophageal sphincter, (b) time required to insert the endoscope past this sphincter, (c) amount of gagging, and (d) name of the physician passing the endoscope. None of these factors, however, differentiated between experimental conditions.

Immediately after completion of the examination, the physician and nurse independently completed the physician-nurse anxiety rating scale. From 2 to 4 hours

after the examination, the Post-Endoscopy Interview Schedule was used to question the patient, and the STAI was completed by the patient.

Results

The results, as a whole, support the prediction that fear during a stressful medical examination would be reduced as a function of the number of prior viewings of a preparation videotape.

Premanipulation Comparisons

Analysis of variance of pretreatment variables revealed that subjects in the three treatment conditions were comparable in age, male-female ratio, reported days of previous hospitalization, and scores on both the Modified R-S scale and the Spielberger A-State scale.

Analyses

Two separate analyses of the data were performed. In the primary set of analyses, the repression-sensitization dimension was ignored and three one-tailed planned comparisons (i.e., E0 vs. E1, E1 vs. E3, and E0 vs. E3) were made for each dependent variable. These were based on the prediction that E3 would evidence the least distress, E1 more distress, and E0 the most distress. Unless otherwise noted, comparisons were made with *t* tests using the appropriate error term from a one-way analysis of variance or, where three repeated measurements were obtained, from a 3×3 analysis (Treatment Condition \times Time Period). The degrees of freedom for the pooled error term, used where repeated measures were obtained, was determined by Satterthwaite's formula (Winer, 1971). A second set of analyses was performed to determine the relationship between coping style and the dependent variables.

Behavioral Measures

Physician-nurse anxiety rating scale. Figure 1 shows the mean anxiety rating by time

¹ *Because I was too Young: The Memoirs of a Distinguished Indiana Physician* can be obtained from the Instructional Media Resource Center, Indiana University School of Medicine, Indianapolis, Indiana 46202.

periods for each of the treatment conditions. The standard deviations for these data ranged from .80 before scoping to .52 after scoping. The E3 subjects were rated significantly less anxious than the E0 subjects both before and during scoping, $t(84) \geq 2.19$, $p < .025$. Group E1 was rated significantly less anxious than E0 during scoping, $t(84) = 1.95$, $p < .05$. The other comparisons did not reach significance.

Tranquilizer required. Diazepam was required by a significantly lower proportion of the E3 patients (20%) than either the E0 patients (45%) or the E1 patients (50%); $z \geq 1.69$, $p < .05$.

Heart Rate

Heart rate during the final minute of the baseline period did not differ between groups, $F(2, 56) < 1.00$.² Heart rate change scores were computed for three time periods by subtracting this basal heart rate from (a) heart rate during the minute immediately prior to insertion of the endoscope, (b) mean heart rate during the first 5 minutes following insertion of the endoscope, and (c) mean heart rate during the second 5 minutes following insertion of the endoscope. The groups did not differ significantly during the minute just prior to insertion of the endoscope. During the first 5 minutes of scoping, Group E3 had significantly less heart rate increase ($M = 18.53$, $SD = 9.39$) than both Group E1 ($M = 27.92$, $SD = 15.84$), $t(101) = 2.64$, $p < .01$, and Group E0 ($M = 24.78$, $SD = 13.58$), $t(101) = 1.80$, $p < .05$. During the second 5 minutes of scoping, the groups did not differ significantly (M s = 16.66, 19.00, and 14.04 for Groups E0, E1, and E3, respectively).

Self-report Measures

STAI. The groups did not differ significantly from each other in state anxiety just prior to the endoscopy. Group means for E0, E1, and E3 were 43.30, 37.80, and 42.25, respectively. Following the endoscopy, the groups were ordered in the predicted direction on state anxiety (M s for E0, E1, and E3 were 39.95, 33.75, and 29.20, respectively), with significance reached for the E0 versus E1 comparison, $t(109) = 1.74$, $p < .05$, and the E0 versus E3 comparison, $t(109) = 3.02$, $p < .005$. As expected the groups did not differ

significantly in trait anxiety after the endoscopy (two-tailed tests). Group means for E0, E1, and E3 were 43.3, 36.0, and 39.5, respectively.

Post-endoscopy interview schedule. The control tape was rated as comparable to the experimental tape in both interest value ($M = 3.7$ for E0, 3.9 for E1, and 4.0 for E3) and the generation of emotional upset ($M = 2.0$ for E0, 2.3 for E1, and 2.0 for E3). Consistent with the habituation/extinction hypothesis, Group E3 rated the third viewing of the experimental tape as less upsetting ($M = 1.5$) than the first viewing of the tape ($M = 2.0$), $t(19)$ for correlated means = 2.78, $p < .005$. The groups were comparable in subject response to the question, "Do you feel that you got too little or too much information about the coming endoscopy?" (M s = 3.1-3.2). No subject indicated that too little information was received.

There were reliable group differences in amount of reported annoyance. To the open-ended question, "Very often some parts of an endoscopy examination cause annoyance. What things annoyed you?" one or more complaints were voiced by nine E0 subjects, eight E1 subjects, and only two E3 subjects. Proportion tests indicated that the number of E3 subjects voicing complaints was significantly lower than the number in either E0 or E1 ($z \geq 2.19$, $p < .025$). Asked to rate how annoyed they felt during the endoscopy, E0 subjects indicated higher annoyance ($M = 1.9$) than either E1 or E3 subjects ($M = 1.4$ for both groups), $t(57) \geq 1.70$, $p < .05$. The groups did not differ in ratings of anxiety experienced during the hour before the endoscopy or in ratings of anxiety and physical discomfort experienced during the endoscopy.

Repression-Sensitization

The dependent variables were also analyzed for differences as a function of subjects'

²One subject was dropped from the heart rate analyses because of failure to obtain a valid basal heart rate. He had the highest basal rate of any patient (111 bpm) and was the only subject to exhibit a drop in heart rate during the stressful endoscopy, demonstrating that a true basal heart rate measurement had not been obtained.

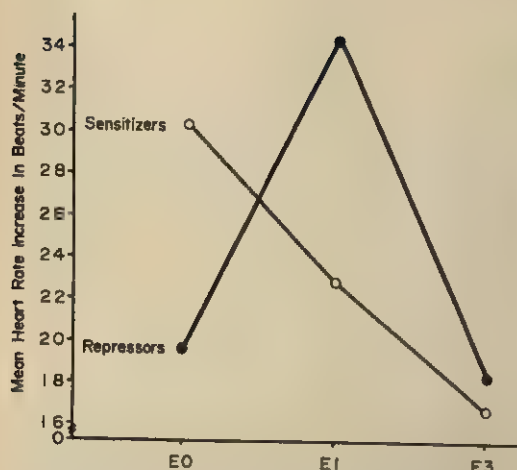


Figure 2. Mean heart rate increase during the first 5 minutes of scoping for repressors and sensitizers in each of the treatment conditions. (E0, E1, and E3 refer to the groups that viewed the control tape once, the experimental tape once, or the experimental tape three times, respectively.)

repression-sensitization score. The 10 subjects whose R-S score fell at the median (12) were eliminated from these analyses, leaving 25 repressors (with scores between 5 and 11) and 25 sensitizers (scores = 13-21). Planned comparisons were used to test the predictions that sensitizers would show a monotonic decrease in anxiety as a function of the number of times they viewed the preparation tape ($E0 > E1 > E3$) and that repressors would show an inverted-U-shaped function, with E1 evidencing the greatest anxiety ($E0 < E1 > E3$). Each comparison was made with a one-tailed *t*-test using the appropriate error term.

Physician-nurse anxiety rating scale. Sensitizers in Group E1 were rated as significantly less anxious than those in Group E0 before, during, and after scoping, $ts(67) \geq 1.75$, $p < .05$. No other comparisons for sensitizers or repressors reached significance.

Tranquilizer required. The percentage of patients in each group requiring diazepam followed the predicted pattern for repressors (% tranquilized = 42, 67, and 9 for E0, E1, and E3, respectively), with the E1 versus E3 comparison reaching significance ($z = 2.58$, $p < .005$). For sensitizers none of the comparisons were significant (% tranquilized = 56, 67, and 43 for E0, E1, and E3, respectively).

Heart rate. There were no significant differ-

ences in heart rate change scores during the minute immediately prior to insertion of the endoscope. Figure 2 provides mean heart rate increases during the first 5 minutes of scoping for repressors and sensitizers in each experimental condition. Sensitizers showed the predicted monotonic decline in heart rate increase, with the E0 versus E3 comparison reaching significance, $t(76) = 2.43$, $p < .01$. Support was also received for the prediction that E1 repressors would have higher heart rate increases than repressors in either E0 or E3, $ts(76) \geq 2.46$, $p < .01$.

A similar pattern was observed during the second 5 minutes of scoping. Sensitizers in E3 showed less heart rate increase ($M = 8.43$) than sensitizers in E0 ($M = 21.04$), $t(76) = 2.27$, $p < .025$. Heart rate increase for sensitizers in E1 ($M = 16.52$) fell between E0 and E3 but did not differ significantly from either of these groups. For repressors, no differences reached significance, but the pattern of results was similar to that observed in the first 5 minutes of scoping. Means for E0, E1, and E3 repressors were 13.31, 21.65, and 16.27, respectively.

State anxiety. There were no differences between groups in state anxiety reported before the endoscopy. After the endoscopy, E1 and E3 sensitizers reported significantly less anxiety than E0 sensitizers, $ts(79) \geq 2.17$, $p < .025$ ($Ms = 44.78$, 33.22, and 29.00 for Groups E0, E1, and E3, respectively). The comparisons on repressors for this time period were not significant.

Post-endoscopy interview schedule. None of the comparisons reached significance for any of the postendoscopy interview items.

Intercorrelations Between Selected Dependent Variables

Table 1 provides Pearson product-moment intercorrelations between the various measures of anxiety and scores on the Modified R-S scale. Intercorrelations between physiological, behavioral, and self-report measures of anxiety were low but frequently significant. Intercorrelations between different self-report measures were moderate. The low nonsignificant correlation between amount of Valium and mean heart rate shows that Valium did not confound the heart rate data.

Table 1
Intercorrelations Between Selected Measures of Anxiety and Repression-Sensitization

	1	2	3	4	5	6	7	8	9	10
1. Mean heart rate change during scoping	—	.28**	.20	.30**	.09	.16	.35**	.44**	.00	.09
2. Total physician-nurse rated anxiety		—	.13	.37**	.31*	.36**	.41**	.42**	.39**	.14
3. Mg of Valium			—	.32**	.11	.31*	.17	.24	.35**	.18
4. State anxiety—baseline				—	.64**	.48**	.53**	.35**	.55**	.29*
5. State anxiety—before endoscopy					—	.49**	.55**	.22	.51**	.34**
6. State anxiety—after endoscopy						—	.30*	.14	.62**	.17
7. Retrospective subject-rated anxiety—before endoscopy							—	.40**	.45**	.16
8. Retrospective subject-rated anxiety—during endoscopy								—	.19	.20
9. Trait anxiety									—	.41**
10. Repression-Sensitization										—

* $p < .05$ (two-tailed).

** $p < .01$ (two-tailed).

Discussion

The results support the hypothesis that fear during a stressful procedure is reduced as a function of the number of prior viewings of a preparation videotape. The ordering of the treatment conditions on the dependent measures was impressively consistent, with E3 generally evidencing the least anxiety, E1 more anxiety, and E0 the most anxiety.

The results are consistent with an extinction/habituation hypothesis. According to this formulation, three exposures to endoscopy-related stimuli resulted in greater habituation and/or extinction of emotion than did one exposure. Likewise, one exposure to the tape produced greater habituation and/or extinction than no exposure. This interpretation is supported by the fact that subjects in the E3 condition rated their third viewing of the tape as less upsetting than their first viewing.

At first glance these results appear similar to those of studies on the vicarious extinction of phobic snake avoidance. Bandura, Blanchard, and Ritter (1969) and Blanchard (1970) found that rated fear and behavioral avoidance decreased as a function of the amount of exposure to filmed models handling a snake. However, in these studies the feared and avoided situation was actually harmless, and the models were pictured as appropriately calm and happy. By contrast, in the present study, the feared situation (endoscopy) was actually aversive and unavoidable. The filmed model

was shown as experiencing discomfort (e.g., gagging, tight muscles, distressed look on face). Since calm behavior was not modeled, it is unlikely that subjects learned to be calmer during the endoscopy through matching their response to that of the model.

The model used in the present study might be labeled a *flooding model* or a *realistically anxious model*, since he showed distress throughout the endoscopy. Several investigators (Kazdin, 1974; Meichenbaum, 1971) have found the viewing of coping models (i.e., those showing initial distress followed by calm contact with the feared object) to be more effective in reducing fear and avoidance than viewing mastery models (i.e., those appearing calm throughout). Theoretically, the sequence of viewing a fearful model become calm facilitates subject identification and imitation. However, the positive findings of the present study using a flooding model suggest that the superiority of results with a coping model may result from the viewing of a fearful model per se rather than from the sequence of viewing a fearful model followed by viewing a calm model. Subject identification with a fearful model can lead to vicarious arousal. When this arousal is maintained long enough and is not reinforced by painful stimulation, it extinguishes. If this were the case, a flooding model would result in greater arousal reduction than a coping model, since the flooding model is fearful throughout the taped sequence rather than only initially. This hypothesis could be

tested by comparing the effectiveness of various amounts of exposure to mastery, coping, and flooding models in producing reductions in arousal.

Another possible explanation for the present findings is provided by the accurate expectancy theory. According to this formulation, the E3 subjects who repeatedly viewed the preparation tape learned more about the endoscopy examination and formed a more accurate expectancy of what was in store for them. This is possible but unlikely, since all groups received detailed verbal information about the examination on at least three separate occasions. The endoscopy examination is relatively simple to understand, and it seems likely that near maximal information was acquired by subjects in all three treatment conditions. In fact, following the endoscopy, the groups were comparable in rated amount of information received, and all subjects indicated they had received enough information. Nevertheless, the experimental design could have been strengthened by administering a test of knowledge about endoscopy just prior to the examination. Andrew (1970) used such a procedure and found no correlation between information learned from an audio preparation-for-surgery tape and her dependent variables.

Another possible explanation for the present results is that the operative variable was simply time spent viewing a videotape, regardless of the content of the tape. However, despite the fact that the control tape was 8 minutes longer than the experimental tape, the E1 subjects evidenced less anxiety than E0 subjects on most of the dependent measures. The results cannot, therefore, be attributed solely to length of viewing time.

Regardless of the theoretical explanation of the results, the finding that anxiety during a stressful medical examination was reduced as a function of the number of prior viewings of a videotape of the examination is of potential practical importance. It suggests that explicit repetitive preparation for stress may be beneficial.

An area of both practical and theoretical concern is the interaction of subject characteristics with the preparation message. In the present study, sensitizers showed the predicted monotonic decrease in anxiety as a function of

the number of videotape viewings. Repressors showed the predicted inverted-U-shaped function for the heart rate and diazepam data, with E1 subjects evidencing the greatest arousal. These findings are consistent with the hypothesis that repressors maintain low arousal in the face of threat by not thinking about it and not seeking information. Theoretically, one exposure to the explicit stimuli of the preparation tape weakened their repressing defenses and left them in an aroused state similar to that of unprepared sensitizers. In E3 the two additional viewings of the tape diminished or extinguished this arousal.

The finding that one viewing of the preparation tape produced increased anxiety in repressors is consistent with the findings of Andrew (1970) and Delong (1971). Delong (1971) found that sensitizers who heard one brief preparation message prior to surgery had less complications and were discharged sooner than sensitizers hearing a control tape. Repressors who heard the preparation tape did not show this positive outcome, and, in fact, they expressed more complaints postoperatively than repressors hearing the control tape. Andrew (1970) also found negative effects for repressors who heard a short audiotape. They required more pain and sleeping medications after surgery.

These results suggest that repressors and sensitizers might benefit from different preparation strategies, with sensitizers prepared extensively and repressors left alone or at least left with their defenses. Repressors might be exposed to a preparation message that supports their defenses by minimizing danger or encouraging avoidance through selective attention. Though they did not differentiate their patient subjects on the repression-sensitization dimension, Langer, Janis, and Wolfer (1975) found beneficial effects from a 20-minute verbal preparation-for-surgery message emphasizing cognitive defensive strategies such as calming self-talk and selective attention. Perhaps this type of preparation message would prove especially effective with repressors.

Taken as a whole, the present study demonstrates that reasonably well-controlled research on preparation for stress can be conducted in real-life stressful situations. This is of particular interest at a time when increased concern for

subjects' welfare and the need for comprehensive informed-consent procedures make the use of analogue populations and stressors less tenable. In addition, findings obtained in a real-life stressful situation seem more likely to generalize to other real-life situations than results obtained in analogue studies that must use relatively low-intensity stressors. Nevertheless, the present findings need to be extended to other settings. This is particularly true of the regression-sensitization findings. As Averill, Olbrich, and Lazarus (1972) noted, findings of relationships between personality variables and stress reactions have tended to "disappear when tested in a slightly different setting" (p. 29).

References

- Andrew, J. M. Recovery from surgery, with and without preparatory instruction, for three coping styles. *Journal of Personality and Social Psychology*, 1970, 15, 223-226.
- Averill, J. R., Olbrich, E., & Lazarus, R. Personality correlates of differential responsiveness to direct and vicarious threat. *Journal of Personality and Social Psychology*, 1972, 21, 25-29.
- Bandura, A., Blanchard, E. B., & Ritter, B. The relative efficacy of desensitization and modeling approaches for inducing behavioral, affective, and attitudinal changes. *Journal of Personality and Social Psychology*, 1969, 13, 173-199.
- Bandura, A., & Menlove, F. L. Factors determining vicarious extinction of avoidance behavior through symbolic modeling. *Journal of Personality and Social Psychology*, 1968, 8, 99-108.
- Blanchard, E. B. Relative contributions of modeling, informational influences, and physical contact in extinction of phobic behavior. *Journal of Abnormal Psychology*, 1970, 76, 55-61.
- Byrne, D. Repression-sensitization as a dimension of personality. In B. A. Maher (Ed.), *Progress in experimental personality research* (Vol. 1). New York: Academic Press, 1964.
- Cassell, S. Effects of brief puppet therapy upon the emotional responses of children undergoing cardiac catheterization. *Journal of Consulting Psychology*, 1965, 29, 1-8.
- Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
- Delong, R. D. Individual differences in patterns of anxiety arousal, stress-relevant information, and recovery from surgery (Doctoral dissertation, University of California, Los Angeles, 1970). *Dissertation Abstracts International*, 1971, 32, 554B.
- Epstein, S., & Fenz, W. D. The detection of areas of emotional stress through variations in perceptual threshold and physiological arousal. *Journal of Experimental Research in Personality*, 1967, 2, 191-199.
- Graham, F. K. Autonomic nervous system habituation. In H. V. S. Peeke & H. J. Herz (Eds.), *Habituation* (Vol. 1): *Behavioral studies*. New York: Academic Press, 1973.
- Hebb, D. O. On the nature of fear. *Psychological Review*, 1946, 53, 259-276.
- Janis, I. *Psychological stress*. New York: Wiley, 1958.
- Johnson, J. E. Effects of accurate expectations about sensations on the sensory and distress components of pain. *Journal of Personality and Social Psychology*, 1973, 27, 261-275.
- Johnson, J. E., & Leventhal, H. Effects of accurate expectations and behavioral instructions on reactions during a noxious medical examination. *Journal of Personality and Social Psychology*, 1974, 29, 710-718.
- Johnson, J. E., Morrissey, J. F., & Leventhal, H. Psychological preparation for an endoscopic examination. *Gastrointestinal Endoscopy*, 1973, 19, 180-183.
- Kazdin, A. E. Covert modeling, model similarity, and reduction of avoidance behavior. *Behavior Therapy*, 1974, 5, 325-340.
- Lader, M. H., & Montagu, J. D. The psycho-galvanic reflex: A pharmacological study of the peripheral mechanism. *Journal of Neurology, Neurosurgery and Psychiatry*, 1962, 25, 126-133.
- Lang, P. J. Fear reduction and fear behavior: Problems in treating a construct. In J. M. Shlien (Ed.), *Research in psychotherapy* (Vol. 3). Washington, D. C.: American Psychological Association, 1968.
- Langer, E. J., Janis, I. L., & Wolfer, J. A. Reduction of psychological stress in surgical patients. *Journal of Experimental Social Psychology*, 1975, 11, 155-165.
- McClelland, D. C. *Personality*. New York: Sloane, 1951.
- Meichenbaum, D. Examination of model characteristics in reducing avoidance behavior. *Journal of Personality and Social Psychology*, 1971, 17, 298-307.
- Melamed, B. G., & Siegel, L. J. Reduction of anxiety in children facing hospitalization and surgery by use of filmed modeling. *Journal of Consulting and Clinical Psychology*, 1975, 43, 511-521.
- Shipley, R. H. Extinction of conditioned fear in rats as a function of several parameters of CS exposure. *Journal of Comparative and Physiological Psychology*, 1974, 87, 699-707.
- Spielberger, C. D., Auerbach, S. M., Wadsworth, A. P., Dunn, T. M., & Taulbee, E. S. Emotional reactions to surgery. *Journal of Consulting and Clinical Psychology*, 1973, 40, 33-38.
- Spielberger, C. D., Gorsuch, R. C., & Lushene, R. E. *Manual for the State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Vernon, D. T. A. Modeling and birth order in responses to painful stimuli. *Journal of Personality and Social Psychology*, 1974, 29, 794-799.
- Vernon, D. T. A., & Bailey, W. C. The use of motion pictures in the psychological preparation of children for induction of anesthesia. *Anesthesiology*, 1974, 40, 68-72.
- Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Addict Descriptions of Therapeutic Community, Multimodality, and Methadone Maintenance Treatment Clients and Staff

Patricia B. Sutker, Albert N. Allain, and Charles J. Smith
Department of Psychiatry and Behavioral Sciences
Medical University of South Carolina

Gary H. Cohen
Tulane University School of Medicine

Adjective Check List (ACL) descriptions of 88 addicts in treatment toward methadone maintenance, multimodality, and therapeutic community clients and program staff within and across rating groups representing the three types of drug treatment conditions were compared. Data analysis procedures included single-groups analyses of variance; combined-groups analyses of covariance with sex, age, race, months addicted, months in treatment, and scores on the Raven Progressive Matrices treated as covariates; and principal-components analysis. Addicts as a group were characterized by high elevations on ACL scales Aggression and Succorance. Client descriptions varied significantly as a function of category rated, with program staff described more positively than client groups, therapeutic community residents described more favorably than other client groups, and methadone clients rated with marked negativity. Between-category differences were most succinctly summarized by factor score comparisons on General Adjustment, one of four factors identified by principal-components analysis and differentially associated with all four categories rated. Results suggest that addict opinions represent a valuable source of ideas for evaluating current treatment approaches and identifying self- and staff perceptions of therapeutic significance.

Treatment successes have been limited among therapeutic approaches to opiate addiction, and pessimism has characterized reports of treatment outcome studies evaluating the effectiveness of strategies such as methadone

This research was supported by General Research Support Grant 22620-GR26 from the Medical University of South Carolina to the first author and was presented at the annual meeting of the Southeastern Psychological Association in New Orleans, Louisiana, March 1976. Special appreciation is expressed to Maggie Pike and Barry Pike of Odyssey House Louisiana and to Ward Suddereth of the Drug Research Clinic who made access to clients in these treatment modalities possible. Appreciation is also expressed to Joanne B. Yoder, Gary S. Gilbert, and John Mehlhorn who assisted in data collection and to Phyllis Goldberg and Larry Libater for instrument scoring.

Requests for reprints should be sent to Patricia B. Sutker, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina, 171 Ashley Avenue, Charleston, South Carolina 29403.

maintenance, multimodality programming, and the therapeutic community. Predicting the long-range fate of addict clients treated in traditional fashion at the U.S. Public Health Service Hospital at Lexington, Vaillant (1970), estimated that 29% of addicts at risk become permanently abstinent each year. Similarly, methadone maintenance programs have been regarded as only partially successful in meeting earlier claims (Dole & Nyswander, 1976), and although Williams and Lee (1975) showed positive behavioral and attitudinal changes among addicts remaining in methadone treatment longer than 3 months, they described a high dropout rate. Other investigators reported that methadone clients continue to use illicit drugs (Chambers & Taylor, 1972) and view methadone as an undesirable solution to their problems (Sutker, Allain, & Moan, 1974). Evaluations of therapeutic communities, however, have been somewhat more positive, and

of those treatment conditions studied over extended time frames the therapeutic community seems to have achieved greatest success (Collier & Hijazi, 1974; Sugarman, 1974). Additionally, short-term personality changes have been associated with residence in therapeutic communities, hospitals, and even prisons (Sutker, Allain, & Cohen, 1974; Zuckerman, Sola, Masterson, & Angelone, 1975).

Although several studies have explored changes in personality, attitudes, or behavior associated with therapeutic intervention, there have been no reports of attempts to collect data in comparable fashion across treatment conditions. Among other difficulties, variability in situational and temporal factors and sampling populations from diverse treatment groups have posed formidable problems for design and have restricted investigators in conducting meaningful comparative research. Additionally, studies attempting to evaluate treatment effectiveness have focused on personality or performance variables, and assessment of the personal attitudes of addict clients themselves toward available treatment approaches, clients, or staff has rarely been an experimental goal. Attitudinal studies have been limited to perceptions of single-treatment modalities (Brown, Bass, Gauvey, & Kozel, 1972; Crowther & Pantleo, 1971), and the consumers of drug treatment services, or addict clients, have been largely overlooked as a source of ideas for systematic evaluation of treatment options.

Assuming that potential or participating clients are information-processing beings whose perceptions of situations provide one key to understanding how they will respond within specific milieus, attitudinal sets may influence client selection and degree of participation within given programs as well as provide valuable information regarding client-perceived successfulness of treatment possibilities. Therefore, the purpose of this investigation was (a) to describe attitudes of addict clients in therapeutic community, multimodality, and methadone maintenance programs toward clients in these conditions and their own program staff and (b) to make comparisons of addict attitudes toward treatment conditions within and across client category groups. It was reasoned that attitudes toward treatment conditions

would best be reflected in measures that required client-rating descriptions of the most obvious products of a treatment program—program clients and staff.

Method

Subject Selection

To sample attitudes toward drug treatment strategies across treatment categories, three programs in New Orleans representing the therapeutic community, methadone maintenance, and multimodality approaches were contacted to request access to program clients. Subsequent to explanation of study purposes and procedures, agreement to participate was obtained from (a) Odyssey House Louisiana, a highly structured therapeutic community with approximately 50 residents, which requires long-term commitment and drug abstinence (client-staff ratio of 6:1); (b) the Drug Research Clinic, a methadone maintenance program with an approximate census of 160, which makes relatively limited demands for individual or group therapeutic involvement (client-staff ratio of 16:1); and (c) the Narcotic Addict Rehabilitation Act Program (NARA), administered by the Department of Psychiatry and Neurology of Tulane University School of Medicine, which offers inpatient treatment and outpatient counseling including methadone maintenance, chemotherapy, and drug-free regimes to approximately 60 clients (client-staff ratio of 6:1).

Each program allowed initial access to clients through individual or group discussions for explanation of the research project. Roughly half of NARA and Drug Research clients were approached following regular appointments, and Odyssey clients were seen in a large group assembled by program staff. Selection was voluntary but with personal solicitation, program encouragement, and promise of \$5 payment. Participation rates in terms of initial willingness varied across programs and ranged from 99% of Odyssey, 95% of NARA, and 70% of Drug Research clients contacted. Of 163 volunteers expressing interest, 75 were eliminated: 11 were unable to read sufficiently well; 61 were unable or unwilling to complete procedures, and 3 had not been enrolled at least 3 months in their specific treatment condition. Dropouts were evenly distributed for NARA and Drug Research programs but were nonexistent among Odyssey residents.

The final sample of 88 treatment addicts included 28 multimodality or NARA clients, 36 therapeutic community or Odyssey residents, and 24 methadone maintenance or Drug Research clients. Racial and sexual composition of treatment programs is presented in Table 1, which reflects an uneven distribution of women, $\chi^2(2) = 6.96$, $p < .05$, and blacks, $\chi^2(2) = 26.92$, $p < .01$, across samples. Distributions were, however, generally representative of breakdowns within sample populations. A summary of subject personal characteristics is outlined in Table 2, and it can be seen that groups differed significantly in age, Raven scores, months addicted to opiates, and months in treatment,

Table 1

Percentage Breakdown by Race and Sex Among Multimodality, Therapeutic Community, and Methadone Maintenance Clients

Treatment group	White	Black	Male	Female
Multimodality—NARA	64	36	93	7
Therapeutic community—Odyssey	83	17	75	25
Methadone maintenance—Drug Research	17	83	62	38

Note. $n = 88$. NARA = Narcotic Addict Rehabilitation Act Program.

which necessitated careful statistical handling of between-group comparisons. All groups shared a long history of opiate addiction, and even the more youthful Odyssey residents were characterized by a mean length of addiction greater than 3 years. Subjects across treatment groups showed a mean of 28 months in drug treatment programs.

Materials and Procedure

Subjects were administered a battery of instruments, which included the Raven Progressive Matrices and four versions of the Adjective Check List (ACL; Gough & Heilbrun, 1965). Used among socially deviant groups in the past (Sutker & Moan, 1973), the Raven was selected to reflect general intellectual or problem-solving ability without undue bias of academic sophistication. The ACL, a self-administered inventory of 300 adjectives yielding scores on 24 need scales, was chosen because it has also been used successfully among addict samples (Brown et al., 1972) and provides a profile of personality attributes. The ACL format was presented four times to subjects who were asked to indicate those adjectives that best described addict clients in a multimodality program, in a methadone maintenance program, in a therapeutic community, and their own program staff. It was reasoned that ratings of clients in the various treatment programs would reflect attitudinal sets and opinions toward the treatment conditions themselves.

Data Analysis

Although treatment groups shared a history of chronic opiate use, criminal activity, and familiarity with drug treatment (over 68%, evenly distributed across sample groups, reported experience with at least two treatment conditions), they differed on several personal characteristic variables as seen in Table 2. In part, differences reflect the complexities of attempting to execute comparative research over varied clinical settings, but sampling to avoid such differences may have rendered samples poorly representative of the groups from which they were drawn. Recognizing the limitations of study implementation, three approaches were pursued in data analysis. Preliminary procedures compared attitudes toward treatment client categories within rating groups by one-way analyses of variance with repeated measures for each ACL scale, and Tukey's tests were performed to evaluate differences for all possible comparisons in which F ratios for categories rated were significant. It was reasoned that extraneous or unplanned between-group differences would in no way contaminate interpretation of these results. Second, attitudes toward treatment client categories were compared within and between rating groups for each ACL scale using Groups \times Categories repeated measures analyses of covariance with sex, race, age, months addicted, months in treatment, and Raven scores included as covariates. The number of independent variables was within suggested limits for sample

Table 2

Group Means and F Values on Subject Characteristic Variables for Multimodality, Therapeutic Community, and Methadone Maintenance Groups

Characteristic	Multimodality	Therapeutic community	Methadone maintenance	F
Age	27.50	23.39	29.79	6.94*
Grade completed	10.36	11.33	10.13	1.87
Raven score	37.07	44.33	26.46	18.98*
Months addicted	71.68	37.39	110.92	11.86*
No. convictions	2.21	1.44	1.00	2.95
Months incarcerated	30.11	10.97	19.00	2.59
Months in treatment	30.64	14.56	43.79	12.08*

Note. $n = 88$.

* $p < .01$.

effects and Tukey's tests were used for significant Groups \times Category interactions. Third, a principal-components analysis with orthogonal rotation was

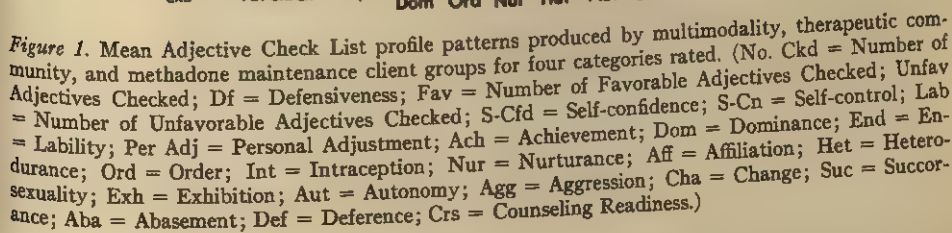


Table 3
Adjusted Mean Adjective Check List (ACL) Values and F Ratios for Program Staff (Staff), Therapeutic Community (TC), Multimodality (Mult), and Methadone Maintenance Clients (Meth) over Treatment Rating Groups (Combined Groups Analyses)

ACL scale	Treatment rating group										F					
	Multimodality (28)					Therapeutic community (36)					Between groups (A)	Between categories (B)	A × B			
	Methadone maintenance (24)															
	Staff	TC	Mult	Meth		Staff	TC	Mult	Meth							
No. Ckd	41.05	42.41	39.87	37.44		40.63	43.08	37.33	37.33	40.20	43.91	43.70	45.62	.56	3.75**	3.44**
Df	48.32	43.10	42.07	27.60		50.54	44.93	28.18	22.26	42.73	40.48	37.19	35.82	2.68	53.95**	9.51**
Fav	46.25	40.61	36.82	22.11		49.08	39.25	22.77	16.86	41.71	38.34	33.42	36.88	3.00	41.54**	8.03**
Unfav	48.86	56.39	56.53	74.68		47.82	53.02	73.24	81.85	60.52	63.52	64.94	69.44	2.27	34.68**	6.12**
S-Ckd	49.13	48.70	47.13	43.34		57.27	51.69	42.41	38.08	46.53	47.86	47.86	45.36	.03	24.41**	11.29**
S-Cn	51.12	45.48	43.65	34.12		48.93	40.10	33.29	30.91	46.84	42.17	42.71	36.88	6.10**	35.59**	2.52*
Lab	41.91	44.13	45.34	41.24		46.55	44.86	44.61	43.38	40.65	41.02	43.98	46.19	.64	.70	2.48*
Per Adj	48.08	40.72	40.36	23.54		47.77	38.33	25.38	19.99	43.09	39.67	38.84	34.63	4.85**	48.81**	6.87**
Arch	49.06	45.70	44.81	34.81		57.01	51.28	35.23	26.84	48.22	46.09	45.17	42.97	1.19	57.64**	15.84**
Dom	50.65	47.97	47.69	38.22		60.07	53.49	37.46	29.71	51.17	49.96	47.42	45.29	1.41	59.96**	16.97**
End	55.17	48.49	46.77	34.92		58.40	50.34	35.03	27.90	47.38	48.34	45.09	41.55	2.98	53.84**	9.84**
Ord	50.50	46.36	45.57	33.64		54.25	46.00	34.88	28.83	47.92	44.59	42.76	40.01	1.98	39.16**	5.59**
Int	48.59	41.84	40.16	27.77		49.10	40.60	28.16	22.18	41.95	41.33	40.41	37.41	3.41*	37.30**	6.95**
Nur	48.77	40.41	40.95	25.27		44.95	39.47	25.00	18.50	39.85	36.18	37.93	33.18	5.27**	37.20**	6.71**
Aff	49.14	43.46	41.53	28.60		46.11	41.08	30.80	24.30	44.02	41.31	39.77	38.02	4.03*	43.63**	5.48**
Het	46.17	42.20	43.88	31.52		45.78	42.67	31.84	26.92	45.59	56.06	55.19	56.81	4.63*	25.83**	6.37**
Exh	52.93	52.36	53.36	54.11		56.43	56.79	53.76	52.85	55.69	50.35	50.52	51.81	1.68	.32	1.23
Aut	43.86	47.61	49.11	51.64		47.54	48.29	51.63	53.35	49.56	50.35	50.52	51.81	1.88	8.30**	.94
Agg	51.05	56.37	58.05	67.69		54.19	58.91	66.89	71.52	58.36	58.61	57.91	63.41	2.72	26.20**	3.35**
Cha	40.25	43.86	43.90	39.68		41.20	46.48	41.78	37.64	42.99	45.74	45.03	47.70	2.46	5.15**	3.13**
Suc	45.91	52.55	51.98	60.98		41.02	49.91	60.77	66.55	50.58	52.49	53.83	58.45	.47	50.14**	7.91**
Aba	48.11	49.78	50.25	51.36		40.35	44.68	50.96	53.60	44.77	46.40	48.19	47.19	2.91	13.27**	4.62**
Def	52.20	47.92	47.42	39.42		41.47	42.58	39.11	37.70	47.05	44.05	46.09	41.68	10.27**	9.97**	1.89
Crs	50.12	53.33	52.16	64.62		48.09	49.06	64.67	65.59	54.68	52.35	53.26	56.89	.94	30.10**	10.04**

Note. Numbers in parentheses are *ns*. Sex, race, age, months addicted to opiates, months in treatment, and Raven score included as covariates. No. Ckd = Number of Adjectives Checked; Df = Degrees of Freedom; Fav = Favorable Adjectives Checked; Unfav = Number of Unfavorable Adjectives Checked; S-Ckd = Self-confidence; S-Cn = Self-control; Lab = Labor; Per Adj = Personal Adjustment; Ach = Achievement; Dom = Dominance; End = Endurance; Ord = Order; Int = Intracession; Nur = Nurture; Aff = Affiliation; Het = Heterosexuality; Exh = Exhibition; Aut = Autonomy; Agg = Aggression; Cha = Change; Suc = Success; Aba = Abasement; Def = Deference; Crs = Counseling Readiness.

* $p < .05$.
** $p < .01$.

performed to identify independent constellations of variables accounting for the major portion of total variance and to provide a concise basis for differentiation of groups in relation to ACL variables.

Results

Within-groups comparisons of client responses across staff and client categories showed that descriptions varied depending on the category rated, and a total of 46 differences were significant at the .05 level of probability. Similar distributions of ACL profile configurations were produced within rating groups, with staff regarded most positively and methadone clients most unfavorably. These relationships are illustrated in Figure 1, which presents mean ACL profile patterns for each category rated within the three treatment groups sampled. Results of within-groups analyses are not reported in greater detail, because mean ACL values differed at most by only tenths of a point from adjusted means derived from combined groups analyses (see Table 3).

Inclusion of sex, race, age, months addicted, months in treatment, and Raven scores as covariates diminished the range of scores by adjusting methadone client ratings in the more positive direction and multimodality and

therapeutic community ratings more negatively. There were, however, differences between and within groups that exceeded the .05 confidence level for statistical significance. Combined groups of treatment clients distinguished among ratings categories in assigning descriptors with significant differences on 22 of 24 ACL scales as seen in Table 3, and inspection of Figure 2 reveals distinct group separation in distribution of ACL configurations. Program staff were described more positively than other groups, followed in order by therapeutic community, multimodality, and methadone clients. Adjectives used to describe methadone clients were most critical and differed from those attributed to therapeutic community residents, multimodality clients, and program staff on 22, 18, and 20 scales, respectively. Ratings of multimodality clients were less favorable than those reported for therapeutic community residents on 17 scales and program staff on 19 scales, and ratings of program staff and therapeutic community residents differed on 13 scales, with staff regarded more favorably.

Treatment groups produced different descriptions of client categories, and Groups \times Categories interaction terms were significant

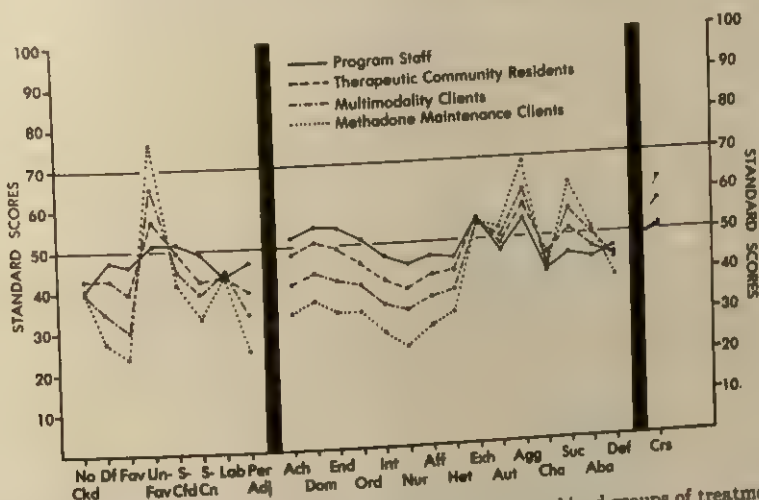


Figure 2. Mean Adjective Check List profile patterns produced by combined groups of treatment clients for four categories rated. (No. Ckd = Number of Adjectives Checked; Df = Defensiveness; Fav = Number of Favorable Adjectives Checked; Unfav = Number of Unfavorable Adjectives Checked; S-Cld = Self-confidence; S-Cn = Self-control; Lab = Lability; Per Adj = Personal Adjustment; Ach = Achievement; Dom = Dominance; End = Endurance; Ord = Order; Int = Intracception; Nur = Nurturance; Aff = Affiliation; Het = Heterosexuality; Exh = Exhibition; Aut = Autonomy; Agg = Aggression; Cha = Change; Suc = Succorance; Aba = Abasement; Def = Deference; Crs = Counseling Readiness.)

for 21 scales (see Table 3). Multimodality clients described themselves and therapeutic community residents similarly but pictured methadone clients more negatively than themselves and therapeutic community residents on 16 scales, including Number of Favorable Adjectives Checked, Number of Unfavorable Adjectives Checked, Self-control, Personal Adjustment, Achievement, Nurturance, and Succorance (see Figure 1). In contrast, they described staff more favorably than themselves on Number of Favorable Adjectives Checked, Self-control, Endurance, Intraception, Affiliation, and Aggression. Methadone clients produced a relatively homogeneous, negative assortment of responses with two significant differences. They described themselves as more succorant and less self-controlled than program staff. Inspection of Figure 1 shows an uncomplimentary choice of descriptors for all categories, with highest elevations on Number of Unfavorable Adjectives Checked, Aggression, Succorance, and Exhibition and lowest scores on Nurturance and Intraception. Therapeutic community descriptions were highly variable, with significant differences on 20 scales, and residents were positive in assessment of their own staff and clients but negative toward methadone and multimodality clients. Program staff, characterized by highest elevations on Self-confidence, Dominance, Achievement, and Endurance and lowest scores on Abasement, Succorance, Change, and Deference, was rated more positively than multimodality and methadone clients on 18 scales each. Therapeutic community residents described themselves less positively than staff on Number of Favorable Adjectives Checked, Self-confidence, Self-control, Personal Adjustment, Achievement, Dominance, Endurance, Order, Intraception, Change, and Succorance scales, but for the most part self-descriptions were in the positive range and followed the pattern attributed to staff with the exception of succorance ratings. Residents rated themselves more positively than multimodality and methadone clients, with significant differences on 20 scales, and they described methadone clients more negatively than multimodality clients on Achievement, Defensiveness, Dominance, Endurance, and Succorance. The latter treatment groups were characterized by prom-

inent elevations on Number of Unfavorable Adjectives Checked, Aggression, and Succorance, with low scores on Number of Favorable Adjectives Checked, Nurturance, Personal Adjustment, Intraception, Affiliation, and Achievement.

Between-groups comparisons of category descriptions showed that multimodality clients were viewed similarly by themselves and methadone clients, but therapeutic community residents attributed more negative adjectives to multimodality clients than did multimodality and methadone groups on 16 scales. Methadone clients were described by therapeutic community and multimodality clients more negatively than they rated themselves on 19 and 13 scales, whereas descriptions of therapeutic community residents tended to be consistent and positive. There were no differences among rating groups in descriptions of therapeutic community residents, with the exception that multimodality clients described residents as higher on Abasement than they viewed themselves. Finally, program staff ratings differed significantly depending on the treatment group. Therapeutic community residents described staff more favorably on Self-confidence, Achievement, Dominance, and Abasement scales than multimodality clients and more favorably than methadone clients described their staff on 8 scales including Self-confidence, Achievement, Dominance, Endurance, and Succorance. Multimodality clients assigned more favorable ratings to staff than did methadone clients on Number of Unfavorable Adjectives Checked, Endurance, and Nurturance.

Principal-components analysis generated four factors with eigenvalues greater than 1.0 as summarized in Table 4. Factor 1, accounting for 51% of the total variance and labeled *General Adjustment*, was characterized by significant loadings (.50 or greater) on 17 of 24 ACL scales including positive loadings for Endurance, Personal Adjustment, Intraception, Nurturance, Defensiveness, Self-control, Affiliation, Number of Favorable Adjectives Checked, and Achievement, and negative loadings for Aggression, Number of Unfavorable Adjectives Checked, and Succorance. Remaining factors accounted for small portions of the variance but are described as follows: Factor 2,

defined by high positive loadings for Autonomy, Exhibition, Aggression, Self-confidence, and Change, appeared to represent Assertiveness; Factor 3, with highest loadings for Abasement, Succorance, and Counseling Readiness, was labeled *Dependence*; and, Factor 4, labeled *Change*, was described by highest loadings for Number of Adjectives Checked, Liability, and Change.

To determine if independent clusters of personality needs represented by the four factors were differentially associated with

Table 4
Orthogonal Rotated Factor Pattern Matrix for
24 Adjective Check List Scales

Scale	Factor			
	1	2	3	4
No. Ckd	.01	-.01	.05	.79
Df	.89	-.03	-.14	.25
Fav	.88	-.10	-.14	.21
Unfav	-.58	.50	.34	.02
S-Cfd	.65	.52	-.30	.05
S-Cn	.89	-.18	.08	-.18
Lab	.23	.37	.15	.56
Per Adj	.92	-.18	-.10	.06
Ach	.86	.21	-.27	.10
Dom	.80	.32	-.41	.07
End	.92	.03	-.20	-.05
Ord	.86	.02	-.15	-.01
Int	.90	-.06	-.09	.10
Nur	.90	-.23	-.05	.05
Aff	.88	-.13	-.01	.27
Het	.77	-.02	-.02	.37
Exh	.12	.83	-.09	.17
Aut	-.29	.84	.03	.01
Agg	-.59	.67	.21	.01
Cha	.15	.51	-.06	.51
Suc	-.55	.27	.66	.24
Aba	-.02	-.13	.91	.13
Def	.71	-.37	.45	-.06
Crs	-.46	.37	.57	-.22
Cumulative proportion of total variance	.51	.66	.75	.79

Note. No. Ckd = Number of Adjectives Checked; Df = Defensiveness; Fav = Number of Favorable Adjectives Checked; Unfav = Number of Unfavorable Adjectives Checked; S-Cfd = Self-confidence; S-Cn = Self-control; Lab = Liability; Per Adj = Personal Adjustment; Ach = Achievement; Dom = Dominance; End = Endurance; Ord = Order; Int = Intrapersonal; Nur = Nurture; Aff = Affiliation; Het = Heterosexuality; Exh = Exhibition; Aut = Autonomy; Agg = Aggression; Cha = Change; Suc = Succorance; Aba = Abasement; Def = Deference; Crs = Counseling Readiness.

client rating groups or categories rated, standardized factor scores generated by principal-components analysis were subjected to univariate analyses of variance and Tukey's tests.¹ Although there were significant differences between and within groups, only differences associated with the categories effect are described in detail, $F(3, 255) = 178.95, 35.23$, and 6.29 for Factors 1, 3, and 4 ($ps < .01$). Constellations of personality needs defined by Factors 1, 3, and 4 were differentially related ($ps < .01$) to the four categories rated. Staff were characterized as higher than client groups on General Adjustment, with therapeutic community residents higher than multimodality and methadone clients and multimodality clients higher than methadone clients. Therapeutic community residents and staff were not described differently on Dependence, but ratings of both groups were significantly less associated with this factor than those for multimodality or methadone clients. Categories rated were also differentially associated with Change, with therapeutic community residents characterized as higher on this dimension than staff and methadone clients.

Discussion

Although treatment addicts shared similar response sets in rating client and staff categories, descriptions varied significantly over categories even with the effects of such potentially powerful variables as age, sex, race, months addicted, months in treatment, and Raven scores partialled out. Differences among category descriptions were most succinctly summarized by comparisons of factor scores on General Adjustment, which was differentially associated with all four categories rated. Staff were described more positively than client groups and were seen as more dominant, better adjusted, less changeable, and less succorant. If staff elicited the most positive descriptors, methadone clients provided the rating stimulus most negatively assessed. Client groups seemed to share a negative set of attitudes toward methadone clients, and by logical extension,

¹ Mean factor scores and F ratios for categories rated within treatment client groups are available on request.

methadone programs, and to evidence greater enthusiasm for multimodality and therapeutic community approaches.

Results are consistent with earlier reports of client pessimism regarding methadone maintenance (Brown et al., 1972; Dole & Nyswander, 1976; Sutker, Allain, & Moan, 1974). Methadone clients were described as more aggressive, more dissatisfied, more hostile and critical, less well-adjusted, less self-confident, and less likely to be dominant, independent, or achievement-oriented than other client groups. Although variations in client composition of treatment programs may have influenced such results, descriptions were relatively uniform across treatment conditions, suggesting that methadone clients were rated without regard for specific programs or client personal characteristics. Addict attitudes toward methadone clients and programs, as well as data from other studies summarized by Cohen, Howard, Klein, and Newfield (1976), do not allow resolution of contradictory viewpoints toward this type of treatment. However, among possible explanations for the nearly unanimous client negativity are perceptions of dubious long-range goals, slow progress toward behavioral change, and less client/staff contact. Whereas other treatment strategies use structure, supervision, and discipline to facilitate development of productive activity, personal independence, and achievement orientation, methadone programs can be seen primarily as mechanisms to sustain dependence on a drug identified with sedative effects and to minimize encounters with law enforcement. The combination of drug dependence and certain daily doses is also one that negates much of the purpose, activity, and excitement previously associated with drug use and client life-style (Preble & Casey, 1969). These speculations suggest the need to determine if methadone programs facilitate development of negative attitudes toward self and others even more than nontreatment, drug-taking conditions.

Clients across treatment categories seemed to hold therapeutic community residents, and perhaps programs, in relatively high esteem. Therapeutic community residents were particularly enthusiastic about themselves and their staff, although they viewed multimodality clients somewhat negatively and methadone

clients with marked disdain. Their variability in descriptor assignments, or response polarization, may be an outgrowth of the totality of program commitment required for continued participation as suggested by the cognitive dissonance model (Festinger, 1957) or evidence that clients strongly endorse program goals and methods. Interestingly, therapeutic community self-descriptions were congruent with the pattern of adjectives assigned to staff, suggesting a high degree of resident/staff identification. Perhaps because of constant interpersonal contact, or association of reinforcement with staff authority, therapeutic community residents aspire more than other groups to mimic their treatment models. The reasons that prompt treatment clients in other conditions to rate therapeutic community residents relatively positively are as yet unknown; however, in view of the extent of personal control relinquished in residential treatment, it is interesting that multimodality clients perceived therapeutic community residents as more inclined toward abasement than they saw themselves.

Despite group differences, clients described themselves in relatively unfavorable terms. Rating patterns across groups showed prominent elevations on Number of Unfavorable Adjectives Checked, Aggression, and Succorance. The results are consistent with reports by Reith, Crockett, and Craig (1975), which point to a constellation of features among addict self-descriptions including exaggerated dependency needs, difficulties in appropriate expression of hostility, excessive demands for interpersonal support and attention, and limitations in responding in kind. Mutual exaggeration of such features provides an important target for therapeutic intervention and assessment of program impact. Although present data do not support the contention that aggressiveness and excessive demands for attention precede initial opiate use, their early identification may signal individuals at high risk. Research exploring personality precursors of drug use and dependence must also take into account interactions with situational as well as more stable environmental, social, and motivational variables.

In summary, addict attitudes represent an interesting and largely untapped source of ideas that may be assessed systematically to

evaluate available treatment approaches and to identify self- and staff perceptions of therapeutic significance. The extent to which positive opinions of staff or fellow clients is related to program participation and ultimate treatment outcome should be explored concomitantly with performance and personality variables across program types in outcome evaluation. Research might also include attitudinal measurement as one methodology to specify treatment parameters perceived as operating within the therapeutic community framework, or conversely, associated with methadone maintenance programs that elicit positive or negative assessments across treatment and nontreatment addict groups and to predict individual performances within a given treatment setting.

References

- Brown, B. S., Bass, U. F., Gauvey, S. K., & Kozel, N. J. Staff and client attitudes toward methadone maintenance. *International Journal of the Addictions*, 1972, 7, 247-255.
- Chambers, C. D., & Taylor, W. J. Patterns of "cheating" among methadone maintenance patients. In W. E. Keup (Ed.), *Drug abuse, current concepts and research*. Springfield, Ill.: Charles C Thomas, 1972.
- Cohen, M., Howard, A., Klein, D. F., & Newfield, K. Evaluating outcome criteria used in methadone maintenance programs. *International Journal of the Addictions*, 1976, 11, 283-294.
- Collier, W. V., & Hijazi, Y. A. A follow-up study of former residents of a therapeutic community. *International Journal of the Addictions*, 1974, 9, 805-826.
- Crowther, B., & Pantleo, P. M. Marathon therapy and changes in attitude toward treatment and behavior ratings. *Mental Hygiene*, 1971, 55, 165-170.
- Dole, V. P., & Nyswander, M. E. Methadone maintenance treatment: A ten-year perspective. *Journal of the American Medical Association*, 1976, 235, 2117-2119.
- Draper, N. R., & Smith, H. *Applied regression analysis*. New York: Wiley, 1966.
- Festinger, L. *A theory of cognitive dissonance*. Evanston, Ill.: Row, Peterson, 1957.
- Gough, H. G., & Heilbrun, A. B. *Manual for the Adjective Check List*. Palo Alto, Calif.: Consulting Psychologists Press, 1965.
- Preble, E., & Casey, J. J. Taking care of business: The heroin user's life on the street. *International Journal of the Addictions*, 1969, 4, 1-24.
- Reith, G., Crockett, D., & Craig, K. Personality characteristics in heroin addicts and nonaddicted prisoners using the Edwards Personality Preference Schedule. *International Journal of the Addictions*, 1975, 10, 97-112.
- Sugarman, B. S. Evaluating drug treatment programs: A review and critique of some studies on programs of the concept house type. *Drug Forum*, 1974, 3, 149-153.
- Sutker, P. B., Allain, A. N., & Cohen, G. H. MMPI indices of personality change following short- and long-term hospitalization of heroin addicts. *Psychological Reports*, 1974, 34, 495-500.
- Sutker, P. B., Allain, A. N., & Moan, C. E. Addict attitudes toward methadone maintenance: A preliminary report. *International Journal of the Addictions*, 1974, 9, 337-343.
- Sutker, P. B., & Moan, C. E. Prediction of socially maladaptive behavior within a state prison system. *Journal of Community Psychology*, 1973, 1, 74-78.
- Vaillant, G. E. The natural history of narcotic drug addiction. *Seminars in Psychiatry*, 1970, 2, 486-498.
- Williams, W. V., & Lee, J. Methadone maintenance: A comparison of methadone treatment subjects and methadone treatment dropouts. *International Journal of the Addictions*, 1975, 10, 599-608.
- Zuckerman, M., Sola, S., Masterson, J., & Angelone, J. V. MMPI patterns in drug abusers before and after treatment in therapeutic communities. *Journal of Consulting and Clinical Psychology*, 1975, 43, 286-296.

Received May 23, 1977 ■

Treatment of General Tension: Subjective and Physiological Effects of Progressive Relaxation

T. D. Borkovec, J. B. Grayson, and K. M. Cooper
University of Iowa

Experiment 1 found virtually no effects of type of no-treatment condition or demand/suggestion on the self-monitoring of daily tension percentage and severity among 43 overly tense college students during a 4-week baseline period. Subjects given four subsequent sessions of progressive relaxation did report significant reductions in tension, which were maintained at a 7-month follow-up. Experiment 2 included 36 overly tense college students and compared a no-treatment condition to groups given nine sessions of relaxation with versus without muscle tension release. Counterdemand instructions were in effect for the first seven sessions. Relaxation with tension release produced reductions in daily tension percentage significantly superior to no-treatment during the counterdemand period, whereas relaxation without tension release did not differ from either group. Treatment effects maintained at a 5-month follow-up. No treatment effects were found on several during-session physiological measures, although Session 1 physiological reduction predicted improvement in tension percentage. Presence or absence of tension release did significantly influence the number of relaxation cycles necessary to produce reports of deep relaxation, frequency of practice, and successfulness of eliminating daily tension at follow-up.

Tension and anxiety represent pervasive adjustment problems. Although these complaints commonly present themselves in outpatient and inpatient centers, the widespread occurrence of tension-related difficulties is aptly exemplified by the 77 million prescriptions for benzodiazepine antianxiety agents (two thirds for diazepam) filled by retail pharmacies in 1972 (Greenblatt & Shader, 1974). Given the cost and potentially hazardous effects of drug intervention, a clear need exists for an effective, nonpharmacological alternative. The present article reports the initial studies in a planned series of investiga-

tions focusing on behavioral intervention strategies for general tension.

Progressive relaxation training (Jacobson, 1938) has an extensive history of clinical application both as a primary treatment strategy for numerous disorders and as a component in the systematic desensitization of phobias (Wolpe, 1958). Even though well-controlled studies in the latter area are common, internally valid designs in the former are infrequent. A review of the literature reveals some cause-and-effect evidence for the progressive relaxation treatment of insomnia (e.g., Nicassio & Bootzin, 1974), childhood asthma (Alexander, 1972; Alexander, Miklich, & Hershkoff, 1972), hypertension (Deabler, Fidel, Dillenkoffer, & Elder, 1973; Shoemaker & Tasto, 1975), tension headaches (Cox, Freundlich, & Meyer, 1975), and phobias (Mathews & Gelder, 1969). Encouraging data from these tension-related problems suggest that relaxation training may be an efficacious procedure for less severe, but more pervasive, daily tension problems in the general population.

Experiment 1 is based on a senior honors thesis conducted by the third author. Both experiments were supported by Grant MH-27484 awarded to the first author from the National Institute of Mental Health and were presented, in part, at the meeting of the Midwestern Psychological Association, Chicago, May 1977.

Requests for reprints should be sent to T. D. Borkovec, Department of Psychology, University of Iowa, Iowa City, Iowa 52242.

Normative Data

A questionnaire administered to college students at the university in the fall and spring of 1975-1976 revealed that 21.2% typically felt tense 50% or more of each day. For subjects from the spring semester ($N = 479$), percentage of daily tension ratings correlated with tension severity ($r = .35$), the social items of Geer's (1965) Fear Survey Schedule ($r = .28$), and Mandler, Mandler, and Uviller's (1958) Autonomic Perception Questionnaire ($r = .27$). Scores on another problem (sleep disturbance) correlated with neither the tension items nor the other measures ($r_s < .07$), suggesting that more than response set to report problems contributed to the relationship between tension ratings and other anxiety-related problems.

Among students for whom tension represents an adjustment problem, specific stressors are readily identified as major contributors to their daily tension. In the first study reported below, subjects filled out daily questionnaires asking them to attribute the day's tension to one or more of six situational factors or to a seventh, "unattributable source," category. Over a 4-week period, the majority of specific tension was attributed to social situations (24.2%), followed by class work (14.5%), tests (14.5%), occupation (4.3%), and public speaking (1.4%); 23.5% was attributed to idiosyncratic sources, whereas the source of the tension was unidentifiable for only 16.7%.

A sizable population of individuals thus exists for whom tension is a daily problem attributable to their student occupation. Prior to committing resources to large-scale outcome studies, we investigated two methodological matters relevant to the use of this target behavior.

Experiment 1

The confounding influence of demand characteristics and suggestions of improvement on measures of behavioral change has become a crucial issue in outcome research (Borkovec, Weerts, & Bernstein, 1977). Unless evidence is obtained that outcome data do not reflect demand/suggestion effects, the internal validity of the study is seriously undermined. The problem is particularly critical for self-report

measures so susceptible to distortion and unreliability. Use of a placebo condition, traditionally regarded as a control for nonspecific effects, has been considered a sufficient methodological device for eliminating the demand confounding. Yet, recent empirical studies indicate that placebos may not establish a level of demand equivalent to that inherent in therapy conditions (cf. Kazdin & Wilcoxon, 1976). By implication, type of no-treatment condition may have a systematic influence on target behavior reports in a similar way. Usual waiting-list conditions may contain subtle demands for improvement during the waiting period or may provide sufficient attention and positive expectancies that actual improvement takes place (Goldstein, 1960). Subjects explicitly informed of their control group status with no mention of future treatment may not be exposed to the same demands or suggestions. In an early attempt to address this issue, Paul (1966) found greater anxiety reduction at follow-up assessment among waiting-list subjects relative to a no-contact control condition. Experiment 1 assessed the effects of demand for improvement and type of no-treatment condition on daily tension reports and provided a pilot evaluation of the effects of relaxation on this problem.

Method

Subjects. Subjects were selected from introductory psychology classes on the basis of two questions on a group testing questionnaire: (a) a 21-point rating scale (0% to 100%) of the percentage of the day they typically felt tense and (b) a 5-point scale (very mild to severely tense) of the severity of that tension. Subjects indicating at least moderate tension for 50% or more of the day were interviewed, given a packet of daily questionnaires, and randomly assigned to four conditions: waiting list or informed control under demand or no-demand instructions for improvement. Of the 53 subjects obtained, 10 were excluded in the final analysis due to voluntary termination, incomplete data, or low baseline severity. Thus, 43 subjects (9 in each waiting-list condition; 11 in the informed-control, demand condition; 14 in informed-control, no-demand condition) completed the study and received course credit for their participation.

Procedure. Subjects filled out a daily tension questionnaire each night before retiring throughout the 6 weeks of the study. The items were identical to the 6 weeks of the study. The items were identical to the group test scales (percent tense and severity ratings). At the time of initial interview, waiting-list subjects were told that 4 weeks of baseline information were first needed. Informed-control subjects were told that

they were in a control group and that their 6 weeks of data would be compared to a therapy group. Within each of these no-treatment groups, demand subjects were informed that research had shown self-monitoring of a problem to result in a reduction of the problem and that they could expect such a reduction to occur in their tension level and severity. Subjects in the no-demand condition were not informed of any potential effects of the self-monitoring.

Waiting-list subjects began treatment after the 4 baseline weeks. Sessions were held either in groups of five or less or individually, depending on scheduling possibilities. Progressive relaxation followed procedures by Bernstein and Borkovec (1973), except that training was presented by a single tape, providing only two tension-release cycles for each muscle group with no provision of additional cycles if tension remained. The first two sessions dealt with 14 muscle groups, and the last two sessions involved a four-muscle-group combination procedure. The tapes were made by the first author. Relaxation was described to the subjects as a skill involving (a) systematic tension release of gross muscle groups to reduce tension and autonomic arousal and (b) learning to identify and relax away extant tension. The importance of daily practice and frequent application was emphasized. After the first session subjects were instructed to practice twice daily (once during a particularly tense time) and to apply the procedure as frequently throughout the day as tension was identified.

Two undergraduate assistants served as training leaders, providing a demonstration of tension-release procedures, answering questions, and starting the taped procedures.

Results and Discussion

Subjects' scores on each of the two questionnaire items (percent tense and severity rating)

were averaged for each of the 4 baseline weeks and the 2 therapy weeks. Table 1 presents the resulting means for each condition.

Baseline phase. A three-way repeated measures analysis of variance (Demand \times No Treatment \times Weeks) on the percent tense measure during the 4-week baseline revealed no effects due to demand or no-treatment factors. The main effect of weeks, $F(3, 117) = 4.49$, $p < .01$, indicated that the percent tense ratings declined during the first 3 weeks (M over weeks = 44.33, 40.05, 36.07, and 37.08). Analysis on the severity data revealed a decline over weeks, $F(3, 117) = 5.01$, $p < .01$ (M s = 2.92, 2.67, 2.62, and 2.59), and a main effect of demand, $F(1, 39) = 5.49$, $p < .03$. Subjects given no demand for improvement reported less tension severity ($M = 2.55$) than did demand subjects ($M = 2.86$).

With one exception, the results indicate that whether or not a subject expects future treatment and whether or not he/she is told to expect improvement have little effect on several weeks of tension ratings. In the one case of a demand effect, the results were opposite to those commonly found in investigations of phobic behavior (Bernstein, 1973). Both percent tense and severity scores declined over weeks. Since spring vacation occurred during the 3rd and 4th weeks, it is unclear whether self-monitoring or a less tension-producing environment was primarily responsible for this decrease.

Table 1

Mean Self-Report Tension Scores for Waiting-List and Informed-Control Groups Under Demand and No-Demand Conditions During Baseline, Treatment Weeks, and Follow-up: Experiment 1

Measure	Group	Condition	Baseline week				Treatment week		Follow-up
			1	2	3	4	5	6	
% tense	Waiting list	Demand	51.4	39.8	40.1	40.9	36.0	30.0	33.2
		No demand	40.2	35.5	29.8	37.4	34.9	26.0	
	Informed control	M	45.8	37.7	34.9	39.2	35.4	28.0	
		Demand	45.2	43.2	39.7	39.5	52.8	54.0	
		No demand	40.5	41.7	34.7	30.4	42.0	37.1	
		M	42.8	42.4	37.2	35.0	47.4	45.5	
	Waiting list	Demand	3.1	2.6	2.7	2.8	2.7	2.5	46.3
		No demand	2.7	2.6	2.3	2.6	2.5	2.3	
Severity	Informed control	M	2.9	2.6	2.5	2.7	2.6	2.4	
		Demand	3.0	2.9	2.9	2.8	3.1	3.0	
		No demand	2.9	2.6	2.6	2.2	2.8	2.6	
		M	2.9	2.7	2.7	2.5	2.9	2.8	

Treatment phase. During treatment, the waiting-list group received relaxation training, and the informed-control group continued as a no-treatment group. With one exception, Treatment \times Demand analyses of variance indicated that the groups did not differ on either dependent measure at the 4th week, the final week prior to relaxation training ($p > .10$). The exception was that no-demand subjects reported less severity than demand subjects, $F(1, 39) = 5.53$, $p < .05$. Three-way repeated measures analyses of variance (Treatment \times Demand \times Weeks) on the 4th, 5th, and 6th weeks found significant Treatment \times Weeks interactions for both the percent tense data, $F(2, 78) = 11.61$, $p < .001$, and the severity data, $F(2, 78) = 8.92$, $p < .001$. On both measures, waiting-list subjects receiving relaxation training continued to decline, whereas informed-control subjects returned to their early baseline level, suggesting that the baseline decline for the total group was due to spring vacation and not to self-monitoring. Subjects given no demand for improvement during baseline continued to report lower levels of tension severity during the therapy period ($M = 2.49$) than demand subjects ($M = 2.82$), $F(1, 39) = 4.67$, $p < .04$.¹

After 7 months, subjects contacted by phone were asked to estimate current, typical percent tense and severity. Although differences were not significant due to large variability, the reports indicated maintenance of improvement among the 14 subjects contacted in the treated, waiting-list condition (M percent tense = 35.8, $SD = 25.3$; M severity = 2.57, $SD = .65$) and maintained baseline levels among the 19 subjects contacted in the informed-control condition (M percent tense = 45.8, $SD = 24.3$; M severity = 2.95, $SD = .91$). The earlier demand condition had no effect on the follow-up reports.

The results from the treatment phase suggest that specific or nonspecific ingredients in relaxation training may provide an effective form of brief intervention for daily tension problems, whereas follow-up indicated maintained improvement for most treated subjects. This conclusion is tentative, since presence or absence of relaxation training was confounded with type of no-treatment condition.

Experiment 2

The second experiment was designed to replicate the relaxation effect on reported tension under more controlled conditions and to initiate efforts to determine the active ingredient(s) within the progressive relaxation procedure.

Although Experiment 1 suggested minimal impact of demand/suggestion on daily tension reports, placebo and Demand \times Treatment interaction effects remained viable hypotheses to explain the outcome improvement. In addition to problems mentioned earlier about the ability of placebo conditions in therapy research to control for client expectancy and demand, the extended use of placebos with suffering individuals raises ethical issues, and alternative control procedures should be explored (O'Leary & Borkovec, in press). One such method, developed in our sleep disturbance program, involves counterdemand instructions: Subjects are told not to expect improvement until after a certain number of training sessions and weeks of practice, since past research has shown that this degree of training and application practice is required before noticeable effects can occur. Statistical comparisons among conditions are then made prior to the end of this counterdemand period. Expectation and demand are hypothetically held neutral during this period, allowing detection of active treatment effects relative to control conditions. The sleep studies have validated this function of the counterdemand procedure (cf. Borkovec & O'Brien, 1976). In Experiment 2, therefore, the counterdemand strategy was used instead of the traditional placebo condition to provide self-report data uninfluenced by main effects of expectancy and demand or their interaction with treatment.

Progressive relaxation consists of two principal procedural components: tension release of gross muscle groups and focused attention on the resulting sensations of tension and relaxation (Bernstein & Borkovec, 1973; Paul, 1966). A 2×2 design involving the

¹ Both dependent measures (percent tense and severity) were also analyzed by Treatment \times Demand \times Weeks analyses of variance with all 6 weeks included. Significant effects identical to those reported above emerged, and no additional effect was found.

presence or absence of each component thus defines the critical comparison conditions for isolation of the active ingredients of the procedure. Over the past 4 years, we have compared progressive relaxation to each component control condition in the treatment of sleep onset disturbance. This set of studies has allowed specific conclusions regarding both the active ingredients of relaxation treatment as well as the maintaining factors of the disturbance itself. Our goal in the present series is to replicate these component comparisons on the general tension problem. Experiment 2 involved the initial comparison between progressive relaxation and relaxation without muscle tension release.

Physiological data were collected during treatment for two reasons: (a) Such data would be useful for identifying the effective mechanism of relaxation treatment of tension, and (b) the study provided an opportunity to carefully assess the physiological effects of progressive relaxation. A review of the existing literature on the physiological effects of relaxation revealed equivocal findings (Borkovec & Sides, in press). Fifteen studies have found relaxation to be superior to control conditions, whereas 10 studies have revealed no differences. The two sets of studies differed significantly on two critical procedural details. Studies demonstrating relaxation superiority involved a greater number of sessions and more frequently used "live" rather than standardized taped instructions. Even though taped training removes a confounding therapist variable, subject control over training progress, an essential aspect of clinical use of the technique, is precluded and may be the reason for the superiority of live training.

In comparing progressive relaxation to relaxation without tension release, therefore, Experiment 2 included nine training sessions (essentially the full course of training recommended by Bernstein & Borkovec, 1973, for clinical use). Second, taped training was used, but by means of two cassettes and subject control over alternation between tapes, subjects controlled their own progress in treatment. Thus, the sessions were identical to the procedure followed in live therapy, but the potentially confounding factors of therapist characteristics and therapist bias were removed.

Finally, typical clinical use of progressive relaxation training involves a progression over sessions from some number of muscle groups to combinations of those muscle groups and ultimately to a recall phase without muscle tension release (Bernstein & Borkovec, 1973). In the latter procedure, the subject presumably learns to quickly and efficiently relax by recalling how the muscle groups felt previously when tensed and released. The clinical assumption of generalization of relaxation over such a progression of increasingly efficient procedures has never been tested. Physiological recordings for progressive relaxation subjects, therefore, occurred during the first training session involving the 14-muscle-group tension-release procedure, during the seventh session involving the 4-group tension release, and during the ninth session involving the 4-group recall procedure. A within-group test of transfer to the recall procedure was thus provided. Second, the inclusion of the relaxation without tension release condition and similar physiological assessments allowed comparisons at Session 7 between tension release and no-tension-release relaxation, and at Session 9, providing a between-groups test of the role of tension release pretraining on relaxation by the recall method. Since the same tape was used for both groups during Sessions 8 and 9, the groups differed only in terms of the presence or absence of tension release training during the initial seven sessions.

Method

Subjects. A new group of 36 undergraduates from introductory psychology were selected on the basis of their responses to three tension items on the group testing questionnaire (50% or greater daily tension, moderately severe or greater, and desire to receive treatment for tension problems). Subjects were randomly assigned within blocks of percent tense levels to three treatment conditions: progressive relaxation with tension release (TR), relaxation without tension release (NTR), and no treatment (NT). Subjects received course credit for participating, but they were not informed of the credit until they made a firm commitment to volunteer for the sake of therapeutic benefit. Although no assessment was made of the degree of their awareness of relaxation techniques, it was assumed that the subjects were relatively naive, since treatment was concluded prior to their introduction to psychotherapy and behavior therapy in their introductory class.

Experimenters. Three graduate students in clinical psychology served as the session leaders and were

counterbalanced across treatment conditions. Two undergraduate assistants served as polygraph technicians and session leaders for make-up sessions.

Dependent measures. All subjects filled out, just prior to retiring at night, a daily questionnaire asking for percent tense ratings, severity ratings, and a self-obtained 60-sec pulse rate. Daily ratings began at least 1 week prior to the first session and ended 1 week after the last session.

Therapy conditions. Subjects in the TR and NTR conditions received nine taped training sessions in their respective procedures over a 5-week period. The NT subjects came to the laboratory for "physiological assessment" sessions temporally corresponding to Sessions 1, 7, and 9 for treated subjects. Subjects in each condition were run in pairs during these recording sessions, whereas treated subjects were trained in groups of six during Sessions 2, 3, 4, 5, 6, and 8. During all training sessions, the group progressed through the procedures at the rate of the slowest subject.

1. Progressive relaxation. Subjects in TR received training in progressive relaxation as described by Bernstein and Borkovec (1973) with two modifications: a) Tensing foot muscles was excluded, and (b) there were only 9 sessions instead of 10. Sessions 1-3 involved 14 muscle groups; Sessions 4-5, a 7-group combination; Sessions 6-7, a 4-group combination; and Sessions 8-9, a 4-group recall procedure.

2. Relaxation without tension release. Relaxation training for NTR subjects was the same as that for TR, except subjects were instructed to identify any tension in each muscle group and to allow those muscles to relax. No actual tensing of muscles occurred, although the identical indirect suggestions of relaxation in TR were provided during each muscle group cycle. Muscle groups and their progressive combinations over sessions were identical to those in TR. During Sessions 8 and 9 (recall method for TR subjects), training tapes were identical for both therapy conditions.

3. No treatment. The NT subjects were told by phone that they would not be able to begin treatment until later in the semester but that they should continue to complete daily questionnaires.

Counterdemand manipulation. Treated subjects were told at the initial contact, after the first session, and daily in a written statement on their questionnaires that past research had shown that they would not notice any treatment effects until after the seventh session. At the eighth session, the session leader told the subjects that if they had been practicing conscientiously, improvement should become noticeable.

Physiological recording sessions. During Sessions 1, 7, and 9, physiological recordings were obtained on a Beckman R611 polygraph situated in an adjoining room. The measures included heart rate, respiration, and electromyogram (EMG) from the frontalis muscles of the forehead.

In Session 1, an introductory tape describing the rationale for the procedure was played, the session leader demonstrated the muscle groups involved, and the electrodes and strain gauges were attached. Two silver-silver chloride electrodes were attached to the

right and left sides of the chest for heart rate recording. Mercury strain gauges were taped across both the subject's chest and abdomen for respiration. Finally, three silver-silver chloride electrodes were attached to the subject's head, one each to the right and left frontal group for EMG and one in the center of the forehead as a ground. Electrode placement was followed by a 10-min adaptation period, during which the subjects were alone in the room and were instructed to lie quietly in recliner chairs with their eyes open. At the end of adaptation, the session leader reentered the room, and the training session began for TR or NTR subjects. NT subjects were asked to relax themselves for another 60 min with the experimenter absent from the room. Relaxation instructions for all nine sessions were presented to TR and NTR subjects via two cassette recorders: one containing instructions for two cycles of each muscle group and one containing repeated cycles, with both recorded by the first author. The session leader turned on the repeated cycles tape whenever a subject signaled remaining tension after two initial cycles, up to a total of four cycles as required. For NTR and TR subjects, 15-sec physiological samples were taken during (a) the last minute of adaptation, (b) the second relaxation cycle of the left biceps, the neck, and the right calf, and (c) for 1 min during the last min of the 2-min quiet period after training was completed. For NT subjects, samples were taken during the last min of adaptation and after the 15th, 30th, and 45th min of the 60-min period and during the last min of the session.

The procedure for Sessions 7 and 9 for TR and NTR subjects was the same, except that four-muscle group training required only about 10 min, and the 15-sec physiological samples were taken during the second relaxation cycle of each of the four-muscle combinations in addition to the adaptation and quiet period samples. For NT subjects the procedure was also similar to their first session, except that they were to relax themselves for only 10 min after the adaptation period, and 15-sec physiological samples were taken at the end of adaptation and at the 2nd, 4th, 6th, 8th, and 10th min of the self-relaxation period.

At the beginning and end of each recording session, subjects completed Husek and Alexander's (1963) Anxiety Differential. At the conclusion of Session 1, treated subjects completed Borkovec and Nau's (1972) therapy credibility questionnaire.

Results

Self-report outcome measures. Each subject's scores on each of the three daily questionnaire items (percent tense, severity rating, and pulse rate) were averaged for the pretherapy week, the week after the seventh session (counterdemand period), and the week after the ninth session (positive demand period). Table 2 presents the resulting means for the three treatment conditions on each self-report measure. One-way analyses of variance found

Table 2

Mean Self-report Tension Scores for Tension Release Relaxation (TR), Relaxation Without Tension Release (NTR), and No Treatment (NT) During Baseline, Counterdemand, and Positive Demand Weeks: Experiment 2

Measure	Condition	Last		
		Baseline	counterdemand	Positive demand
% tense	TR	52.6	40.0	35.8
	NTR	42.1	33.4	25.4
	NT	44.5	48.5	45.4
Severity	TR	3.0	2.8	2.4
	NTR	2.7	2.5	2.0
	NT	2.9	3.0	3.0
Pulse rate	TR	76.7	75.9	74.4
	NTR	75.5	72.1	72.9
	NT	67.7	68.7	70.1

no differences among the groups during the baseline week on the percent tense and severity data. A significant treatment effect was found on the pulse rate measure, $F(2, 31) = 4.37$, $p < .05$. Scheffé post hoc comparisons (Hays, 1963) indicated that neither TR nor NTR differed significantly from NT, although NT was significantly lower than the treated groups combined.

1. Counterdemand period. Two-way repeated measures analyses of variance (Treatment \times Weeks) were conducted on each questionnaire item from baseline to the seventh session. The Treatment \times Weeks interaction was significant only for the percent tense measure, $F(2, 31) = 5.12$, $p < .02$. Scheffé post hoc comparisons revealed that the combination of TR and NTR showed significantly greater reductions in tension than NT, whereas TR alone was significantly superior to NT; NTR did not differ from TR or NT. Covariance analysis applied to the pulse rate scores indicated no treatment effect.

2. Positive demand period. The same analyses applied to each measure from the baseline week to the week following the ninth session found significant Treatment \times Weeks interactions on all three measures, percent tense, $F(2, 31) = 6.69$, $p < .004$; severity, $F(2, 31) = 15.22$, $p < .001$; pulse rate, $F(2, 31) = 3.56$, $p < .05$. Covariance analysis of the pulse rate data revealed no treatment effect. Scheffé post hoc comparisons on the percent tense and severity measures indicated (a) no differences between TR and NTR and that (b) the two

treated groups, separately and combined, displayed significantly greater reductions than NT.

3. Frequency of practice, frequency of application, and treatment credibility. Three additional daily questionnaire items were completed by treated subjects once treatment was initiated: (a) frequency of formal relaxation practice during the day, (b) frequency of identifying daily tension and of attempting to apply relaxation to eliminate that tension, and (c) frequency of successful elimination of identified tension. Weekly averages of each measure and of the ratio of successes to attempted applications were submitted to a Treatment \times Weeks repeated measures analysis of variance. The main effect of treatment was significant on frequency of practicing, $F(1, 20) = 8.87$, $p < .008$. The practice frequency of TR subjects was close to the recommended twice-a-day schedule ($M = 1.78$), whereas that of NTR subjects was lower ($M = 1.43$). No effect emerged from analysis of the frequency of application, whereas the frequency of successful application increased significantly over weeks among the treated subjects, $F(6, 120) = 3.79$, $p < .002$. The ratio of successes to attempts also increased over weeks, $F(6, 120) = 13.89$, $p < .001$, with success incrementing from 44.7% after the first therapy session to 81.8% after the ninth session.

Credibility questionnaires (Borkovec & Nau, 1972) administered after the first session and after the final week of the study indicated that

the TR and NTR groups did not differ on reported confidence in their respective procedures. Correlations between credibility scores and improvement on the percent tense and severity measures during both counterdemand and positive demand periods were all nonsignificant.

Self-report process measures. Two-way repeated measures analyses of variance (Treatment \times Session) were applied to the Anxiety Differential scores obtained at the beginning and end of recording sessions (1, 7, and 9) and to the presession-postsession difference in Anxiety Differential scores. Only the main effect of weeks was significant on the presession measure, $F(2, 62) = 4.94, p < .02$. Subjects reported less anxiety at the beginning of the seventh and ninth sessions than at the beginning of the first session ($M_s = 71.1, 64.9$, and 66.0). On the postsession measure, the main effect of treatment was significant, $F(2, 31) = 5.09, p < .02$. The no-treatment group displayed the highest anxiety level at the end of the sessions, followed by TR and NTR ($M_s = 64.4, 59.2$, and 52.3 , respectively). None of the Scheffé pairwise comparisons were significant. Finally, the main effect of sessions was significant on the presession-postsession difference scores, $F(2, 62) = 3.26, p < .05$. The greatest decrements in anxiety occurred from the beginning to the end of the first session ($M_s = -11.9, -7.3$, and -6.5 for sessions 1, 7, and 9, respectively).

During treatment sessions, subjects in TR and NTR received two relaxation cycles on each muscle group before proceeding to the next group. Additional cycles were administered to any subject who signaled remaining tension. Over all nine sessions, the total number of additional cycles required to achieve reports of complete relaxation was 8.5 for the TR condition and 17.8 for the NTR condition, $t(22) = 2.10, p < .05$.

Physiological process measures. Physiological data were amplified on the Beckman polygraph, fed into a Hewlett-Packard FM tape recorder, and then reduced by a PDP-12 computer. Five respiration measures were derived from each sample: sec/cycle, sec/inspiration, sec/expiration, sec/inspiration to sec/expiration, and inspiration amplitude/expiration amplitude. Two cardiac measures

were obtained: heart rate in beats/min and heart rate variability. Too few EMG responses were recorded to allow meaningful analysis.

One-way analyses of variance conducted on the Session 1 adaptation period revealed no differences on any measure among the three treatment conditions prior to therapy. Three-way repeated measures analyses of variance were then performed. In addition to the treatment and session (1, 7, and 9) factors, these analyses included a phase factor with five levels (an adaptation sample, three muscle group samples, and a quiet period sample). The three muscle group samples had been obtained in Session 1 at the left biceps, neck, and right calf. Sessions 7 and 9, however, involved four muscle group combinations with samples taken after each group, (a) hands, forearms, and biceps; (b) face and neck; (c) chest and abdomen; and (d) thighs and calves. Consequently, the average of the latter two major muscle groups was taken for Sessions 7 and 9 to provide the third muscle group sample.

The only significant effect involving the treatment factor to emerge from analysis of the five respiration measures was a Treatment \times Phase interaction on the inspiration/expiration amplitude ratio, $F(8, 120) = 2.57, p < .05$. Inspection of Table 3 means indicates that inspiration amplitude, relative to expiration amplitude, increased for TR and NTR conditions from baseline to the first or second muscle group samples and then declined to near-adaptation levels. The NT condition

Table 3
Inspiration/Expiration Amplitude Means for Treatment Groups During the Five Phases of the Sessions

Phase	Group		
	TR	NTR	NT
End of adaptation			
1	1.00	1.02	1.03
Muscle group sample			
2	1.08	1.04	.97
3	1.06	1.05	1.08
4	1.01	1.03	1.02
Quiet Period			
5	1.01	1.03	1.01

Note. TR = tension release; NTR = relaxation without tension release; NT = no treatment.

showed a marked reduction in the ratio at the time corresponding to the first muscle group sample for treated subjects and a marked increase at the second sample before declining again below adaptation level.

No effects involving the treatment factor emerged from analysis of the heart rate data. Heart rate did decline from adaptation to quiet period for the total group ($M_s = 67.4$ to 65.0 , respectively), $F(4, 108) = 4.29$, $p < .01$. Heart rate variability declined initially and then rose during the five phases ($M_s = 4.9$, 4.0 , 4.7 , 4.7 , and 5.4), $F(4, 64) = 2.59$, $p < .05$.²

Correlations between physiological change and subjective improvement. To further assess the relationship between physiological activity and subjective tension, correlations for the total group of subjects were computed between physiological change during Sessions 1 and 7 and improvement in percent tense and severity scores during the counterdemand period (Sessions 1-7). Physiological change scores for each of the seven measures were obtained from the difference between the quiet period sample and the adaptation sample. Counterdemand improvement on percent tense was significantly related to four of the seven Session 1 change scores (sec/cycle, $r = -.57$; sec/expiration, $r = -.51$; sec/inspiration, $r = -.48$; heart rate variability, $r = -.36$). Thus, reduction in respiratory functioning and increased heart rate variability during Session 1 predicted outcome improvement on the percent tense measure. Only one correlation was significant between percent tense change and physiological change during Session 7 (sec/inspiration to sec/expiration, $r = .39$). Counterdemand improvement on the severity measure correlated significantly only with one Session 1 physiological measure (sec/cycle, $r = -.33$) and was not related to Session 7 measures.

Physiological relationships with improvement during the positive demand period were of little interest, since demand and expectancy effects may have contributed to outcome reports at that time. The fact that only one of the seven physiological measures (Session 1 heart rate variability) correlated significantly with improvement in percent tense ($r = -.39$) and severity ($r = -.42$) suggests that those positive demand outcome measures were indeed a function of nonphysiological variables.

Finally, although the self-reported pulse rate measure failed to reflect treatment effects due to pretherapy differences among the treatment groups, its potential usefulness in future research was supported by the finding that pretherapy pulse rate reports correlated significantly with continuously monitored heart rate during the adaptation period of Session 1, $r(26) = .46$, $p < .02$.

Follow-up contact. Ten TR and 7 NTR subjects were located 5 months later and phoned by a research assistant "blind" to their condition status. Subjects estimated current daily tension percentage, rated the severity of current tension and the effectiveness of their relaxation skill in reducing daily tension on 5-point scales, and indicated how often they used their relaxation skills during the 5-month period. Average percent tense scores were 24.0 for the TR condition and 20.3 for the NTR condition, indicating maintenance of the gains reported immediately at the end of treatment. Average severity ratings fell between "mild" and "moderate" levels of tension (2.70 for TR and 2.29 for NTR), representing slight though nonsignificant increases in severity since the end of treatment. The two treated groups did not differ significantly on either measure. On the third item, TR subjects reported being "quite" successful ($M = 4.10$, $SD = .74$) in applying the technique to eliminate tension, whereas the average success rating of NTR subjects fell between "very little" and "somewhat" ($M = 2.71$, $SD = 1.98$). Although the mean difference only approached significance (Mann-Whitney $U = 1.60$, $p < .12$), the variances of the two groups were significantly different, $F(6, 9) = 7.13$. Finally, TR subjects reported still practicing relaxation 2.11 times per week ($SD = 2.01$), whereas NTR subjects practiced only .55 times ($SD = .72$). The variance test was significant, $F(9, 6) = 7.77$, $p < .01$.

² Recordings from NT subjects during the first session were obtained for 60 minutes to guarantee that self-relaxation occurred at least as long as the TR and NTR sessions. Average duration of the first session for the two latter conditions was 36.0 minutes. All Session 1 physiological measures were therefore reanalyzed comparing TR and NTR to NT separately during its 30-minute, 45-minute, and 60-minute samples. No significant treatment effects were found for any measure at any of the three NT time samples.

whereas a Mann-Whitney test indicated that the group means also differed significantly ($z = 2.41, p < .02$).²

Discussion

The results of Experiment 2 lead more firmly to the conclusion that progressive relaxation can produce significant reductions in the daily tension experienced by overly tense individuals. The percent tense improvement was greater for treated subjects than for untreated controls. The fact that this difference was obtained during a counterdemand period suggests that some active ingredient within progressive relaxation procedures contributed to the subjective improvement that occurred, independent of demand, expectancy, and/or their interaction with treatment. Follow-up reports at 5 months indicated that these gains were maintained over a fairly long interval subsequent to treatment.

Although progressive relaxation differed significantly from no treatment during counterdemand, relaxation without tension release fell between the two latter conditions and differed significantly from neither. The critical procedural component in progressive relaxation was thus not unambiguously identified. Apparently, frequent attempts to relax while focusing on internal sensations are sufficient to promote tension reduction. Additional results, however, suggest a potentially important role for muscle tension release. Subjects in progressive relaxation required significantly fewer training cycles to produce subjective reports of complete relaxation. This result suggests that tension release relaxation may be a more efficient procedure and perhaps explains the absence of differences between the two treatment conditions on the during-session Anxiety Differential. Furthermore, progressive relaxation subjects practiced their procedure significantly more often both during treatment and over the succeeding 5 months.

Although it is unclear which aspect of tension release procedures accounts for these results, frequent practice of an adaptive skill is obviously desirable. Finally, perhaps due to the additional practice, progressive relaxation subjects tended to report greater success, with significantly less variability, in eliminating

daily tension 5 months after treatment. All of these differences, due to the presence or absence of muscle tension release, occurred between conditions that did not differ in credibility or the amount of improvement expected.

Our review of the literature had suggested that the present study maximized the probability of obtaining significant physiological reduction effects: several sessions, subject controlled treatment, and a presenting problem with some significance for daily adjustment. The absence of treatment effects on physiological process leads to two separate observations. First, the main difference between our procedures and those of previous studies finding significant relaxation effects (Borkovec & Sides, in press) was the presence or absence of a therapist. We are forced to conclude that the therapist factor may be a critical variable in promoting physiological reduction during relaxation training. Second, the fact that some relationships were found between physiological process and subjective improvement suggests that the subject's ability to reduce physiological activity by any procedure, including self-relaxation, does contribute to reductions in subjective tension. However, the treatment effects on the subjective outcome measures indicate that such improvement will occur only if the subject realizes that relaxation is a skill to be practiced and applied to his/her daily life.

Hopefully, future experiments involving comparisons of progressive relaxation with other component control conditions will further elaborate the mechanisms by which relaxation promotes reductions in daily tension. With such specifications, it should be possible to develop increasingly efficient and efficacious procedures for ameliorating a pervasive adjustment problem.

² After the conclusion of the treatment period of the study, NT subjects were given similar progressive relaxation training with three exceptions: (a) No physiological recording was conducted, (b) no counterdemand instructions were administered, and (c) subjects were not required to complete daily questionnaires during their treatment period. With these differences in mind, 5-month follow-up information obtained on 5 NT subjects indicated mean values generally similar to those of the TR subjects (percent tense = 27.5, severity = 2.80; effectiveness of application = 3.25; and frequency of practice = 1.45).

References

- Alexander, A. B. Systematic relaxation and flow rates in asthmatic children. Relationship to emotional precipitants and anxiety. *Journal of Psychosomatic Research*, 1972, 16, 405-410.
- Alexander, A. B., Miklich, D. R., & Hershkoff, H. The immediate effects of systematic relaxation training on peak expiratory flow rates in asthmatic children. *Psychosomatic Medicine*, 1972, 34, 338-394.
- Bernstein, D. A. Behavioral fear assessment: Anxiety or artifact? In H. Adams & P. Unikel (Eds.), *Issues and trends in behavior therapy*. Springfield, Ill.: Charles C Thomas, 1973.
- Bernstein, D. A., & Borkovec, T. D. *Progressive relaxation training*. Champaign, Ill.: Research Press, 1973.
- Borkovec, T. D., & Nau, S. D. Credibility of analogue therapy rationales. *Journal of Behavior Therapy and Experimental Psychiatry*, 1972, 3, 257-260.
- Borkovec, T. D., & O'Brien, G. T. Methodological and target behavior issues in analogue therapy outcome research. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification*. New York: Academic Press, 1976.
- Borkovec, T. D., & Sides, J. K. Critical procedural variables related to the physiological effects of progressive relaxation: A review. *Behavior Therapy*, in press.
- Borkovec, T. D., Weerts, T. C., & Bernstein, D. A. Behavioral assessment of anxiety. In A. Ciminero, K. Calhoun, & H. Adams (Eds.), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Cox, D. J., Freundlich, A., & Meyer, R. G. Differential effectiveness of electromyograph feedback, verbal relaxation instructions, and medication placebo with tension headaches. *Journal of Consulting and Clinical Psychology*, 1975, 43, 892-898.
- Deabler, H., Fidel, E., Dillenkoffer, R., & Elder, S. T. The use of relaxation and hypnosis in lowering high blood pressure. *American Journal of Clinical Hypnosis*, 1973, 16, 75-83.
- Geer, J. H. The development of a scale to measure fear. *Behaviour Research and Therapy*, 1965, 3, 45-53.
- Goldstein, A. P. Patient's expectancies and nonspecific therapy as a basis for (un)spontaneous remission. *Journal of Clinical Psychology*, 1960, 16, 399-403.
- Greenblatt, D. J., & Shader, R. I. Drug therapy: Benzodiazepines. *New England Journal of Medicine*, 1974, 291, 1011-1015.
- Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.
- Husek, T. R., & Alexander, S. The effectiveness of the Anxiety Differential in examination stress situations. *Educational and Psychological Measurement*, 1963, 23, 309-318.
- Jacobson, E. *Progressive relaxation*. Chicago: University of Chicago Press, 1938.
- Kazdin, A. E., & Wilcoxon, L. A. Systematic desensitization and nonspecific treatment effects: A methodological evaluation. *Psychological Bulletin*, 1976, 83, 729-773.
- Mandler, G., Mandler, J. M., & Uviller, E. T. Autonomic feedback: The perception of autonomic activity. *Journal of Abnormal and Social Psychology*, 1958, 56, 367-373.
- Mathews, A. M., & Gelder, M. G. Psychophysiological investigations of brief relaxation training. *Journal of Psychosomatic Research*, 1969, 13, 1-12.
- Nicassio, P., & Bootzin, R. A comparison of progressive relaxation and autogenic training as treatments for insomnia. *Journal of Abnormal Psychology*, 1974, 83, 253-260.
- O'Leary, K. D., & Borkovec, T. D. Conceptual, methodological, and ethical problems of placebo groups in psychotherapy research. *American Psychologist*, in press.
- Paul, G. L. *Insight vs. desensitization in psychotherapy*. Stanford, Calif.: Stanford University Press, 1966.
- Shoemaker, J., & Tasto, D. Effects of muscle relaxation on blood pressure of essential hypertensives. *Behaviour Research and Therapy*, 1975, 13, 29-43.
- Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford, Calif.: Stanford University Press, 1958.

Received June 2, 1977 ■

Behavioral Contingencies and Self-mutilation in Lesch-Nyhan Disease

Lowell Anderson, Joseph Dancis, and Murray Alpert
New York University School of Medicine

Lesch-Nyhan syndrome is a rare, sex-linked, recessive disease that is accompanied by severe self-mutilation, especially finger biting. Evidence is presented suggesting that parental response patterns may contribute to the genesis of the self-injurious behavior (SIB). The therapeutic effectiveness of punishment, positive reinforcement of either SIB or non-SIB, and time-out learning paradigms were evaluated. Electric skin shock failed to suppress the behavior. Positive reinforcement of non-self-injury and time-out from social reinforcement were consistently and rapidly effective, indicating a complex interaction of genetic and environmental factors in the production of SIB. Elimination or major reductions in incidence of SIB was maintained during follow-up periods of 2 years.

Lesch-Nyhan disease, which is characterized by hyperuricemia, stunting of growth, choreoathetosis, generalized spasticity, and compulsive self-mutilation, is a rare sex-linked genetic disease (Berman, Balis, & Dancis, 1969). The genetic mutation is responsible for a deficiency of hypoxanthine phosphoribosyltransferase activity, an enzyme involved in purine metabolism (Rubin, Dancis, Yip, Nowinski, & Balis, 1971; Seegmiller, Rosenbloom, & Kelley, 1967). Although mental retardation was originally thought to be a necessary concomitant of Lesch-Nyhan disease (Nyhan, 1968), another study described a child with the disease who had normal receptive intelligence (Scherzer & Olson, 1969). However, the children are severely impaired in motor function and are unable to walk or sit without support. Speech is dysarthric and difficult to understand.

The diagnosis is suggested by the symptomatology associated with excessive urinary excretion of uric acid and is confirmed by

demonstrating the specific enzyme defect in red or white blood cells or skin fibroblasts. Family patterns of inheritance and mosaicism in tissues of the obligatory heterozygote (Migeon, Der Kaloustian, & Nyhan, 1968; Silver, Cox, Balis, & Dancis, 1972) indicate an X-linked recessive disease. Management strategies are supportive of kidney function by preventing excessive production of uric acid with allopurinol, thereby greatly increasing life expectancy. Questions concerning procedures for managing the behavioral consequences of the illness and for optimizing the life adjustment of these children thus become relevant.

One of the most striking and disabling characteristics of the disease is a pattern of self-injurious behavior (SIB) that affects virtually all Lesch-Nyhan children (Lesch & Nyhan, 1964). Near the end of the second year of life, they begin to bite their fingers, lips, or arms with sufficient force to cause lacerations, expose tendons, and even amputate fingertips. If not prevented, SIB occurs at a very high rate, accompanied by general agitation and crying. Reliance on tubular arm splint restraints to prevent finger biting and extraction of teeth to prevent lip biting have constituted the only therapeutic approach to SIB in Lesch-Nyhan children.

Although no direct studies have been reported that demonstrate normal pain sensi-

This research was performed in the Clinical Research Unit of New York University School of Medicine in Bellevue Hospital and was supported by National Institute of Health Grant HDO 4526.

Requests for reprints should be sent to Lowell T. Anderson, Department of Psychiatry, New York University School of Medicine, 550 First Avenue, New York, New York 10016.

tivity, observers agree that pain thresholds do not appear to be elevated in Lesch-Nyhan children and that the children are manifestly distressed by the tissue damage. In fact, it is commonly observed that the children appear placid and passive while in restraints but become agitated and distressed if the restraints are removed. It also seems clear that the pattern of SIB in these children differs from the pattern of occasional and accidental injury sometimes reported in individuals with congenital lack of pain sensitivity.

In our preliminary investigations, using naturalistic observations and careful interviews with parents, a pattern of possible environmental support for the behavior began to emerge. The frequency of self-injury varied considerably and predictably with variations in the environment. For example, we recorded attempts at SIB in one child while he was cared for by his grandfather, his mother, and his grandmother. During a 15-minute period when the grandfather cared for the child, there were 110 attempts to self-injure on the first day of observation and 104 on another day. With the mother, the corresponding figures were 74 and 77, whereas with the grandmother there were "only" 10 and 12 attempts at self-injury. The chronological order of these observations was counterbalanced to exclude an artifact of a time trend.

The consistency of different response rates in the presence of different family members and the wide range of response rates among family members suggest a behavior that is under specific discrimination. Thus, although it is clear that the SIB is intimately related to the genetic defect, important environmental contributions may also be presumed. A series of studies were therefore initiated to identify the contingencies that supported SIB and to design an intervention to reduce the use of restraints.

Method

Subjects

Five males, C.S., P.B., C.W., R.K., and J.A., 3, 5, 11, 12, and 13 years old, respectively, were studied while they were inpatients in the clinical research unit of New York University School of Medicine. All cases had been previously studied for extended periods, and the diagnosis had been confirmed by demonstration of

the enzymatic defect. The oldest patient was institutionalized, whereas the four younger ones lived at home. They demonstrated the characteristic motor retardation and spasticity and could not sit or stand without support. Prior to treatment the children wore arm restraints and bore scars on fingers and lips. Upon removal of restraints, they would immediately become agitated and begin finger biting with an intensity that would lead to severe injury in a matter of minutes. In addition to finger biting, other SIB such as lip biting, head banging, or neck snapping were present in all children.

Despite the fact that all were severely impaired in verbal expression and motor control, it was our impression that the receptive intelligence of these five children ranged from mildly retarded through normal. Because of their extremely poor verbal and motor abilities, attempts to administer intelligence tests were either not undertaken or produced scores that seemed far too low to be a valid indication of IQ. However, the children are responsive to social situations and to speech. Three of the children can count, tell time, and have memorized the television schedule. At times, they seem to speak in sentences, although none can read. The unusual nature of the disease makes it difficult to accurately evaluate their intelligence, but they are not severely retarded.

Written informed consent was obtained from both parents of all children, and our procedures were approved by the Medical Center's Human Research Committee. Parents and therapists experienced the shock levels delivered to the children.

Procedure

Four learning paradigms were investigated in all children—punishment, positive reinforcement of non-SIB by noncontingent attention, positive reinforcement of SIB by contingent attention, and time-out contingent on self-injury. These paradigms were compared to a response prevention baseline. Overriding clinical considerations prevented us from adhering to a rigid research design. Number of learning trials, sessions, and order of investigation of learning contingencies varied slightly from child to child. The specific information provided in Figure 1. Throughout the study, two independent raters recorded the number of attempted SIBs. Observer reliability was found to be good (Pearson $r = .94$).

Response prevention. The child was seated in a wheelchair, and his arms were released from the restraint. An assistant, seated in front of the child, intercepted the child's hand as it approached his mouth and placed it back in the child's lap without speaking or otherwise interacting with the child.

Punishment. Subjects C.S., P.B., C.W., and R.K. were given electric finger shock contingent on each hand-to-mouth contact. The fifth child, J.A., was given shock contingent on attempts at head banging. Shock was delivered by a Scientific Prototype isolated shock source (102K) through electrodes attached to the fingers of the hand that was most frequently bitten. In all cases the shock intensity was 3 mA, a level generally described as quite painful but well below that which

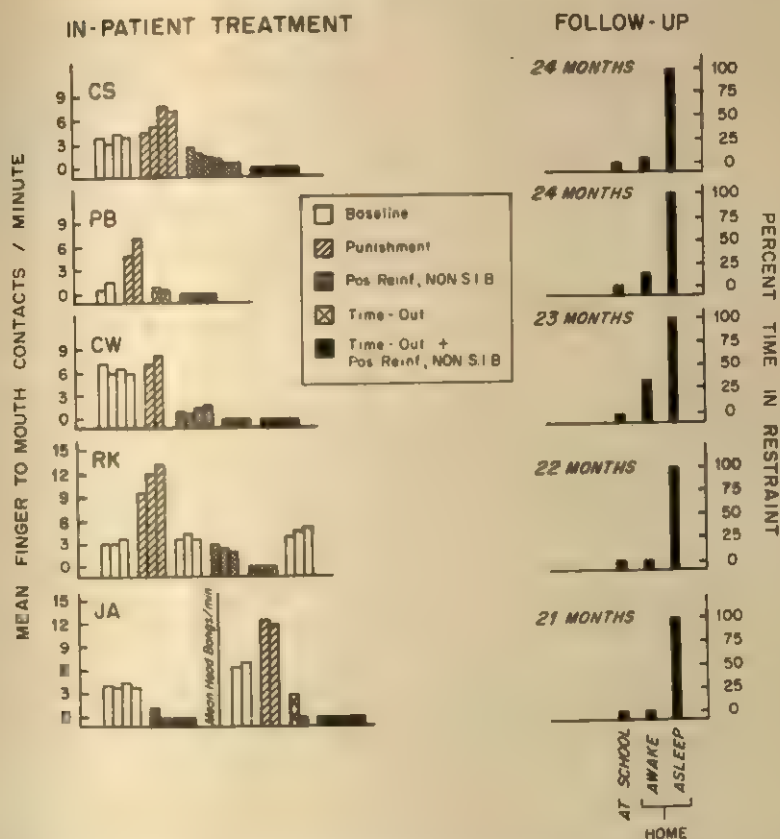


Figure 1. Self-injurious behavior (SIB) during therapeutic intervention and at follow-up. (Left side of chart indicates frequency of SIB during 30-minute treatment sessions administered by therapist. Right side of chart records time spent in restraints at home and in school, at follow-up. Pos. Reinf. = positive reinforcement. CS, PB, CW, RK, and JA are the initials of the subjects.)

reduces tissue damage. The response prevention procedure was used concurrently with finger shock.

Positive reinforcement of non-SIB by noncontingent attention. When the child was not agitated or attempted to bite, the therapist would smile, talk, play, or in other ways interact with the child.

Reinforcement consisted of verbal reinforcement ("good boy," "you're doing fine," etc.) and stroking the head, patting the head, and so forth. The first session began with the removal of one arm restraint. The therapist used the response prevention procedure to prevent injury and would hold and stroke the arm and maintain conversation only during periods of calm. Usually, reinforcement was given for very brief periods of calm behavior and gradually developed into longer intervals. After biting had been eliminated for intervals of several minutes, the second arm restraint was removed.

Positive reinforcement of SIB by contingent attention. The purpose of this approach was to evaluate, under laboratory conditions, the effect of the management commonly used by parents in the home. The therapist responded to each attempted SIB with a quick intervention to prevent injury while making reassuring

statements and stroking the child. Non-SIBs were ignored.

Time-out contingent on SIB. Time-out consisted of the withdrawal of all attention to the child following each SIB attempt. As the hand began moving toward the mouth, the therapist simply turned away from the child for about 5 sec. The response prevention procedure of intercepting hand-to-mouth contacts was discontinued.

Results

Figure 1 presents the rate of SIB attempts by each child for each of the treatment modalities in which that child participated. Each bar represents the data from three $\frac{1}{2}$ -hour sessions. Usually three sessions were conducted each day. Note the relatively constant level of SIB under baseline conditions. In all cases punishment failed to suppress the frequency of SIB and, in fact, it appeared to act as a facilitator.

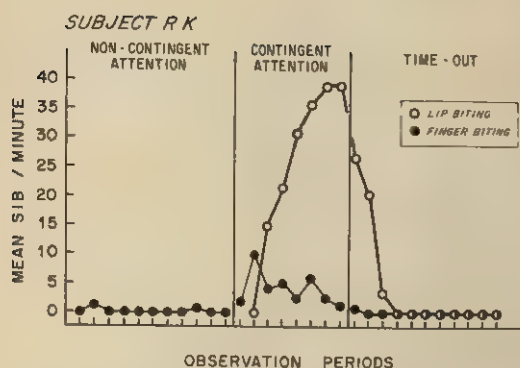


Figure 2. The effect of noncontingent attention, contingent attention, and time-out frequency of self-injurious behavior (SIB) in 12-year-old Lesch-Nyhan child R.K. (Each point represents the mean hand-to-mouth contact per minute in a 5-minute observation period.)

The rate of self-injury ranged from 5 to 14 per minute, making the total number of learning trials for each 30-minute "punishment" session substantial.

Positive reinforcement for non-SIB produced a decrease in the attempts for Subjects C.S. and C.W. but not for R.K. The procedure was not tried with Subjects P.B. and J.A. Time-out was associated with a decrease in SIB for Subjects P.B., C.W., R.K., and J.A. but was not tried with C.S. With the exception of Subject R.K., for whom the procedure was not used, the combination of time-out with positive reinforcement of non-SIB eliminated self-injury in all cases.

The results of positive reinforcement of SIB by contingent attention are presented in Figure 2. This procedure was used with the 12-year-old child (Subject R.K.) and was conducted prior to therapeutic training. On Day 1, the child was given 1 hour of noncontingent attention (reinforcement of non-SIB), 40 minutes of attention contingent on self-injury (reinforcement of SIB), and 1 hour of time-out. On Day 2, the order of the time-out and noncontingent attention procedure was reversed. For each of the 2 days, the sessions were continuous with no pause between manipulations. Each data point represents the mean number of finger- or lip-biting responses per minute for blocks of 5-minute observations averaged over the 2 test days. Both noncontingent attention (reinforcement of other behaviors) and time-out contingent on self-injury resulted in virtual elimination of self-injury. Positive reinforcement of SIB by giving attention contingent on self-injury, however, produced a rapid and dramatic increase in the rate of finger and lip biting.

An extinction curve can be observed that extends into the early minutes of the time-out period (following termination of the 40-minute period of contingent attention). The reinforcement by contingent attention period, originally planned for 1 hour, could not be completed because the SIB occurred with increased rate and force and the therapist was unable to respond quickly enough to prevent damage.

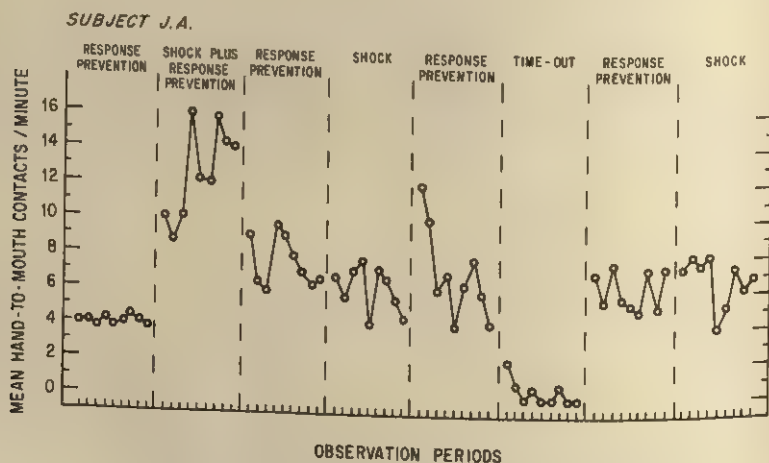


Figure 3. Effects on self-injurious behavior for Subject J.A.

Figure 3 contains a summary of the results during procedures intended to differentiate the influence of positive reinforcement by contingent attention intrinsic to response prevention from the effect of shock. Prior to therapeutic training, response prevention alone, response prevention plus finger shock (3 mA), finger shock alone, and time-out were investigated in Subject J.A. Four 15-minute observations were made each day for 6 consecutive days. Each manipulation was studied during three of these observation periods. Response prevention was used as a baseline against which the effects of the other procedures were assessed. Thus, the response prevention baseline procedure both preceded and followed each of the other manipulations. This child was selected for this investigation because his teeth had been extracted and it was possible to eliminate the response prevention procedure without fear of serious damage. Each data point represents the rate of SIB per 5-minute period.

The response prevention procedure was evaluated four times during the 6-day test. On each occasion there was a consistent and relatively stable rate of between four and seven attempts at self-injury per minute. When shock was administered without interrupting the behavior (response prevention discontinued), the self-injury rate remained at about the level of the response prevention "baseline." When shock and response prevention were combined, the rate increased to almost twice that of either procedure alone. Time-out produced a rapid decrease in SIB to less than one per minute.

Generalization Training

Following elimination of SIB by time-out and positive reinforcement of non-SIB behavior, additional therapists were introduced in sequence. In all cases, by the fifth therapist, the generalization process was complete and SIB was eliminated in the presence of all hospital personnel. Parents were then brought to the hospital and trained in the use of time-out and positive reinforcement of non-SIB procedures. The parents received 1-3 days of training. The training was then transferred to the home, and, depending on the ease of

training, the therapist visited the home for 3-7 days.

Follow-up

The follow-up period ranged from 22 to 24 months. During this time, a therapist was available, on request, for home visits to analyze difficulties and refine the therapeutic skills of the parents. Rarely, however, did the therapist directly intervene in a behavioral problem, limiting herself to advice and instruction. Most requests for additional assistance were not for recurrences of SIBs but for problems such as temper tantrums, breath holding, spitting, cursing, or other antisocial or noncompliant behaviors that had not been treated in the hospital.

For a 1-week period at the end of follow-up, each family was asked to record the daily frequency of SIBs during mealtimes, bedtime, and while bathing or changing diapers. The parents also recorded the total amount of time their child spent in and out of restraints and the conditions surrounding the use of restraints.

The youngest child (C.S.) was without arm restraints for 18 months following discharge, at which time the parents decided to use them occasionally during high-risk periods (defined as when the child was "cranky," tired, or during automobile rides). It appeared that restraints were easier for the parents than the constant vigil required to maintain appropriate contingencies. Most recurrences of self-injury could be attributed to a specific event such as an untrained grandmother moving into the house or the illness of the mother necessitating a radical change in the daily routines.

SIB has been almost completely absent in the school setting. The rating form revealed that on the average, the child was in restraints 5% of the day. No incidents of self-injury were noted during the rating period, although the mother used the time-out procedure approximately three times per day to interrupt threatened finger biting. At bedtime the child was placed in restraints.

The 5-year-old subject (PB) remained free of arm restraints in school. At other times restraints were used under specific circumstances such as the bus ride to and from school. The behavior was under a high degree of

stimulus control. With certain adults or specific settings, SIB did not occur. During the 1-week period when the rating form was in use, the child was in restraints 12% of the day, and self-injury in the form of biting the inside of the mouth occurred, on the average, twice per day. At bedtime the child was placed in restraints.

The 11-year-old subject (CW) has remained free of restraints while at school and at bedtime. At home he remained free of restraints for 6 months following treatment, at which time he was placed again in arm braces by his parents. The failure to maintain the gains in the home coincided with the development of a chaotic situation following the hospitalization of the mother and the inability of the father to care adequately for the child. One finger-biting event was recorded in the final week of follow-up, but there were numerous instances of head snapping. He was in restraints approximately 39% of the day.

The 12-year-old subject (RK) remained completely free of restraints during the day but was placed in elbow braces at bedtime. This child has never engaged in SIB during the night, and the restraints seemed more related to parental habit than necessity. No SIBs were recorded during the 1-week period when the rating form was in use.

The oldest boy (JA), who is institutionalized at a state school, has been without restraints for the entire follow-up period. His teacher and a nurse's aide with primary-care responsibilities for the child have responded independently to the questionnaire and to personal interview. Neither have reported the need of either day or night restraints and neither are aware of any recent attempts at self-injury. A physical examination of the child did not reveal any signs of recent self-injury.

In summary, there has been a considerable reduction in the dependence on restraints for all of the children. With two children, arm restraints have been discontinued entirely during the day. For another two, the restraints are applied at infrequent intervals, usually in response to the parent's wish to temporarily end the need for maintaining appropriate contingencies rather than to treat SIB. The fifth child, after 6 months free of restraints, is now in arm splints almost half of the day. This

is undoubtedly related to a turbulent home situation, with both parents being invalids and with the child having unruly siblings. Most important to educational and social development, the children are usually free of arm restraints in school. At night three of the five children are placed in restraints.

Discussion

To conduct an operant conditioning analysis without permitting self-injury, a response prevention procedure was instituted. Following removal of the restraints, a therapist seated in front of the child interrupted movement of finger to mouth just before injury could be inflicted. The frequency of such movements provided the baseline against which to measure the effectiveness of a variety of interventions.

Because of the reported effectiveness of the punishment paradigm in the alleviation of SIB associated with other diseases (Lovaas & Simmons, 1969), electric shock administered to the fingers was the first conditioning modality that was tried. In most instances, there was an increase in finger-to-mouth movements (Figure 1). The increased rate of SIB attempts was attributed to the effect of shock superimposed on the response prevention procedure. Separate analysis of these two components was possible in J.A., whose teeth had been previously removed to prevent finger biting. Shock and response prevention were roughly equivalent in their ability to maintain the rate of SIB (Figure 3). The unexpected ineffectiveness of aversive stimulation contrasted with the rapid and consistent success in eliminating self-mutilation using time-out and positive reinforcement of non-SIB. There did not appear to be any replacement of SIB with other undesirable behaviors. Spitting and cursing were common, but these behaviors usually antedated treatment.

In other diagnostic categories, punishment has been found to be consistently and rapidly effective in reducing the rate of SIB (Butcher, & Lovaas, 1968; Corte, Wolfe, & Locke, 1971; Lovaas & Simmons, 1969; Risley, 1968; Tate, 1972; Tate & Baroff, 1966; Young & Wincze, 1974). Although failures to maintain the gains produced by contingent electric skin shock frequently occur (Harris & Romanczyk, 1976;

Jones, Simmons, & Frankel, 1974; Romanczyk & Goren, 1975), no report could be located that indicated a total ineffectiveness of the procedure. The failure of contingent shock to reduce the rate of self-injury thus seems to be peculiar to Lesch-Nyhan disease and may therefore be a clue to the action of the genetic defect and how it interacts with environmental experience to produce the behaviors characteristic of this disease.

Significant gains, as measured by freedom from arm restraints, were maintained in all children during an observation period of 2 years. The opportunity to attend school without restraints permitted participation in educational and social activities, making for happier children and easing tensions at home. The gains at home were also significant but were more irregular, indicating that either teachers can more reliably maintain the appropriate contingencies or perhaps a busy teacher, nurse's aide, or an unsuspecting stranger are less likely to attend to finger biting, thus creating a discriminative period for non-self-injury.

The maintenance of long-term improvement in general is better than that reported for SIB associated with autism, schizophrenia, or mental retardation (Harris & Romanczyk, 1976; Romanczyk & Goren, 1975). This may be related to the presumed higher intelligence of the Lesch-Nyhan child and the ability to participate in socially appropriate activities once freed of physical restraints. This ability to interact with their environment when not restrained, and the loss of that activity contingent on self-injury, may provide the necessary contingencies to account for the persistence of treatment effects.

It is evident from this study that the SIB of Lesch-Nyhan disease is strongly influenced by environmental factors. Contingent attention of SIB serves as a strong reinforcer (see Figure 2). Inexperienced attendants, such as parents and grandparents, are likely to react to SIB with contingent attention. On the other hand, when trained in time-out and appropriate reinforcement techniques, the same personnel can maintain elimination of the undesirable behavior. Although SIB must be associated in some way with the primary enzymatic defect of Lesch-Nyhan disease, a

deficiency in hypoxanthine phosphoribosyltransferase, it is not an inevitable consequence. The sequence of events leading to the establishment of SIB is not known, but its maintenance is dependent on environmental support.

In selecting children for this treatment, a careful analysis of the home situation and parental motivation is essential. Parents should be carefully trained and candidly warned of the difficulties they will encounter, and initiation of treatment should be postponed until the parents are able to make use of consultation. In training parents or therapists to administer this therapy, there are three treatment issues to be emphasized.

First and most important, in the early stages of training, time-out or isolation should be used with caution and only under the direction of a skilled therapist. Time-out is an extinction procedure, and therefore the schedule of reinforcement that has maintained the behavior is critically related to the rate of behavior change. Because of the constant vigilance that these children require, most SIB is maintained by almost continuous reinforcement. Behavior maintained on a high-density reinforcement schedule is the most amenable to extinction procedures, and it is important that this advantage not be lost. Care should be taken that self-injury is not reinforced during time-out or isolation, thereby increasing its resistance to extinction and accidentally rendering the time-out or isolation procedures ineffective. If parents or therapists have already trained a child to self-injure in the absence of direct visual contact, an abrupt change in the salient characteristics of the isolation situation or a gradual errorless learning procedure may prove helpful. It would be unwise to initiate therapy unless isolation can be made a discriminative period for non-self-injury.

Second, in designing continuing care procedures, it should be remembered that self-injury has proven to be an attention-eliciting behavior. The likelihood of maintaining good treatment effects depends heavily on the child's ability to gain attention in socially appropriate ways despite the extreme limitations on physical movement imposed by the spasticity. Self-help skills such as eating without assistance, arranging access to television, books, the out-of-doors, and so on, must be devised. Our

most successful cases had willing and capable parents as well as access to educational institutions in which training in self-help skills was emphasized.

Third, the results of this study suggest that Lesch-Nyhan children are either reinforced by, do not learn from, or are unable to inhibit a response associated with a "punishment" contingency. This effect is specific to punishment procedures in that the children can learn to inhibit a response when a time-out contingency is used. From the parents' point of view, it appears that any attempt to use discipline or punishment for any behavior (on close scrutiny, the range of inappropriate behavior is endless) is not only ineffective but frequently has the effect of increasing the rate of behavior. Inasmuch as discipline is a primary means of socialization and parental reliance on the technique is so pervasive, it is unreasonable to expect that the therapist's instructions can be adequately followed on a daily basis. Thus the procedures described above should be viewed as management techniques for specific behaviors (e.g., finger or lip biting) occurring in specific situations (e.g., physical therapy or other formal educational experiences) and not as a general cure for the manifold behavioral problems associated with Lesch-Nyhan disease.

References

- Anderson, L., Dancis, J., Alpert, M., & Hermann, L. Punishment, learning and self-mutilation in Lesch-Nyhan disease. *Nature*, 1977, 265, 461-463.
- Berman, P. H., Balis, M. E., & Dancis, J. Congenital hyperuricemia, an inborn error or purine metabolism associated with psychomotor retardation, athetosis and self-mutilation. *Archives of Neurology*, 1969, 20, 44.
- Butcher, B., & Lovaas, O. I. Use of aversive stimulation in behavior modification. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior 1967: Aversive stimulation*. Coral Gables, Fla.: University of Miami Press, 1968.
- Corte, H. E., Wolfe, M. M., & Locke, B. J. A comparison of procedures for eliminating self-injurious behavior of retarded adolescents. *Journal of Applied Behavior Analysis*, 1971, 4, 201-213.
- Harris, S. L., & Romanczyk, R. C. Treating self-injurious behavior of a retarded child by over-correction. *Behavior Therapy*, 1976, 7, 235-239.
- Jones, F. H., Simmons, J. Q., & Frankel, F. An extinction procedure for eliminating self-destructive behavior in a 9-year-old autistic girl. *Journal of Autism and Childhood Schizophrenia*, 1974, 4, 241-250.
- Lesch, M., & Nyhan, W. L. A familial disorder of uric acid metabolism and central nervous system function. *American Journal of Medicine*, 1964, 36, 561-570.
- Lovaas, O. I., & Simmons, J. Q. Manipulation of self-destruction in three retarded children. *Journal of Applied Behavior Analysis*, 1969, 2, 143-157.
- Migeon, B. R., Der Kaloustian, V. M., & Nyhan, W. L. X-linked hypoxanthine guanine phosphoribosyltransferase deficiency: heterozygote has two clonal populations. *Science*, 1968, 160, 425-427.
- Nyhan, W. L. Summary of clinical features in Lesch-Nyhan syndrome. *Federation Proceedings*, 1968, 27, 1034-1041.
- Risley, T. R. The effects and side effects of punishing the autistic behaviors of a deviant child. *Journal of Applied Behavior Analysis*, 1968, 1, 21-34.
- Romanczyk, R., & Goren, E. Severe self-injurious behavior: The problem of clinical control. *Journal of Consulting and Clinical Psychology*, 1975, 43, 730-739.
- Rubin, C. S., Dancis, J., Yip, L. C., Nowinski, R. C., & Balis, M. E. Purification of IMP: Pyrophosphate phosphoribosyltransferases, catalytically incompetent enzymes in Lesch-Nyhan disease. *Proceedings of the National Academy of Sciences*, 1971, 68, 1461.
- Scherzer, A. L., & Ilson, J. B. Normal intelligence in the Lesch-Nyhan syndrome. *Pediatrics*, 1969, 44, 116-120.
- Seegmiller, J. E., Rosenbloom, F. M., & Kelley, W. N. Enzyme defect associated with a sex-linked human neurological disorder and excessive purine synthesis. *Science*, 1967, 155, 1682-1684.
- Silver, D. N., Cox, R. P., Balis, M. E., & Dancis, J. Detection of the heterozygote in Lesch-Nyhan disease by hair root analysis. *New England Journal of Medicine*, 1972, 286, 390.
- Tate, B. G. Case study: Control of chronic self-injurious behavior by conditioning procedures. *Behavior Therapy*, 1972, 3, 72-83.
- Tate, B. G., & Baroff, G. S. Aversive control of self-injurious behavior in a psychotic boy. *Behaviour Research and Therapy*, 1966, 4, 281-287.
- Young, J. A., & Wincze, J. P. The effects of the reinforcement of compatible and incompatible alternative behaviors on the self-injurious and related behaviors of a profoundly retarded female adult. *Behavior Therapy*, 1974, 5, 614-623.

Received August 2, 1977 ■

On the Decision to Be Assertive

Decky Fiedler
Pacific Lutheran University

Lee Roy Beach
University of Washington

This study examines the applicability of an expectancy/decision model to assertiveness in a nonclinical population. Assertiveness, defined here as refusal to comply with an unreasonable request, has been researched extensively from the viewpoint of behavior theory, which prescribes anxiety reduction and skill acquisition for the training of assertive behaviors. However, little has been done to investigate the reasons why assertive behavior occurs in one situation and not in another. The results of this study suggest that participants, irrespective of their scores on standard measures of assertiveness and of anxiety, consider the consequences of being assertive when making a decision about how to behave. Moreover, it was found that the difference between participants who chose an assertive response and those who did not lies in the formers' assessments of the probabilities that bad consequences will occur and good consequences will not rather than in their evaluations of how bad or how good those consequences would be. These results imply that training programs should take into account the participant's perceptions of the risks involved in being assertive and that the focus should be on changing these perceptions rather than on attempting to change his or her values or focusing solely on specific assertive behaviors.

Studies of assertiveness and designs for assertiveness training programs have revolved around two hypotheses about why some persons tend to be less assertive than they might. One hypothesis (Wolpe, 1958, 1969; Wolpe & Lazarus, 1966) is that nonadaptive anxiety inhibits the expression of assertiveness. The second hypothesis (Lazarus, 1971) is that unassertive people lack the necessary skills for assertiveness. In practice, these hypotheses are not mutually exclusive and may occur simultaneously (Wolpe & Lazarus, 1966), but they are presented as such to highlight the most salient features of assertive behavior as described in each theory. To complement and extend these views, we propose a third hypothesis: Prior to acting

either assertively or nonassertively, people weigh the consequences that could be expected to result from either behavior and they elect the behavior that appears most favorable. That is, the decision to act assertively is not a general trait. Instead, it varies in any situation according to the consequences expected by the person involved. Differences between persons who tend in general to be assertive and those who tend in general to be less so lie in differences in their expectations about these consequences.

Assertiveness training programs are designed to help people who have problems with interpersonal communication as a result of overly passive or overly aggressive behaviors. Using techniques such as instruction, modeling, rehearsal, and feedback, training attempts to reduce the anxiety of interpersonal encounters and to teach specific behaviors such as appropriate eye contact, appropriate voice tone, use of "I" statements, appropriate body language, and so on. These programs have been successfully offered to a variety of clinical populations (e.g., Hersen, Eisler, & Miller, 1973).

This research was partially supported by the Office of Naval Research Contract N00014-76-C-0193 (Terence R. Mitchell and Lee Roy Beach, principal investigators).

The authors wish to thank Martha Perry, David Hathaway, and Diana Stacy, Robyn Hembree, and Lynn Bowen.

Requests for reprints should be sent to Decky Fiedler, Department of Psychology, Pacific Lutheran University, Tacoma, Washington 98447.

Even though there is evidence that as a result of decreased anxiety and acquired assertiveness skills, there are changes in both assertive behavior and self-report of assertiveness (McFall & Lillisand, 1971; Rathus, 1973), there are still theoretical difficulties. One difficulty is that if one adopts a very extreme interpretation of either of the original models, training that lowers anxiety or teaches skills should result in the trainee being rather consistently assertive in all subsequent situations that call for assertiveness. This, however, is patently not the case. One problem is that behaviors encouraged by training often are not rewarded, or are even punished, by others. A woman who is learning to be assertive may find that she was more highly valued when she was accommodating, self-denying, and quiet. Her assertiveness may increase her self-respect, but she may be unwilling to accept the reactions of others to her behavior and therefore cease to use her new skills. Even though in actual practice assertiveness trainers tacitly recognize this difficulty, neither a pure anxiety reduction theory nor a pure response acquisition theory approach can account for a subject's own evaluation of his or her response.

The second difficulty lies with the "lack of skills" hypothesis: There is evidence that unassertive persons *do* know the appropriate assertive responses. Gottman and Schwartz (Note 1) found that nonassertive persons did not differ from highly assertive persons in their ability to construct a written response or verbally deliver an assertive response in a hypothetical situation. The difference between the two groups of persons did not become apparent until they were confronted with a situation that was highly similar to an actual interpersonal confrontation. In short, nonassertive people may well know what to do, but in stressful or high-risk situations they tend not to do it.

Although both skill acquisition and anxiety reduction models explain some of the consequent behavior change experienced during assertiveness training, a cognitive approach can add to our understanding of this process by exploring the conditions under which a person will choose to act in an assertive manner and how that person

then evaluates his or her response. We propose therefore, that a person's willingness to be assertive, defined here as his or her refusal to comply with an unreasonable request, can be predicted from knowing how he or she evaluates the possible consequences of being assertive in various kinds of situations. What differentiates assertive and nonassertive persons is less a matter of their "personalities" or differences in their repertoire of skills than it is of differences in how they evaluate the consequences of being assertive in different kinds of situations. This hypothesis implies that if a person maintains a rather stable, say, negative, evaluation of a specific class of situations, it would be expected that he or she would characteristically be nonassertive in such situations but not necessarily in other situations. It is also likely to be true that in any interaction a person may evaluate the consequences of assertion to be negative due to the specific characteristics of the situation (who is involved, the timing of the response, past history of response with that individual, etc.). In a similar interaction with different topographical attributes, assertion may be perceived as the desirable response.

Decision/Expectancy Theory

Decision-expectancy theory provides a model for describing choice behavior. The basic assumptions are that people act in a fairly rational way and that their behavior is determined by their expectation that the behavior will lead to various consequences and their evaluation of these consequences (Mitchell & Biglan, 1971). Choosing the action that will result in the maximum expected long-range gain requires comparing the alternatives in terms of the decision maker's utility for each of the possible consequences and in terms of the decision maker's expectation that the action will result in the attainment of each of these outcomes; the action that promises the larger expected gain (or the least expected loss) is the one to be selected.

Actually, there are two similar but formally different models. The first, expectancy theory, as described by Fishbein and Ajzen (1975), is concerned with the relationships among

beliefs and values and behavioral intent (BI). BI is defined as the intention to perform a behavior, and its magnitude is assessed by obtaining the person's subjective judgment about whether or not he or she will perform that behavior. The attitude toward a behavior, and therefore BI, is related to the person's evaluation of potential consequences, measured on a 7-point bipolar affective dimension, and a similarly measured belief about the likelihoods that the consequences will occur. In the present case, a subject's evaluation of various consequences that could result from assertion would be multiplied by his or her belief that these will occur as a result of assertion. The sum of these products is added to a component that reflects relevant social pressures (usually obtained from ratings about the opinions of significant others, social norms, etc.), and this sum is assumed to be monotonically related to the magnitude of the person's intention to perform (BI) the assertive act in question.

The second model, subjective expected utility (SEU), comes from the area of decision theory (Edwards, 1954). It differs from the expectancy model primarily in its greater mathematical simplicity, but the logic is quite similar; when the person must decide between an assertive and a nonassertive course of action, he or she must evaluate these actions in terms of the utilities of their positive consequences (U), the utilities of their negative consequences (\bar{U}), and the subjective probabilities (P and $1 - P$) of these consequences occurring if one or the other course of action is selected.

$$\begin{aligned} SEU_A &= P_{1A}U_1 + P_{2A}U_2 + \dots + P_{nA}U_n \\ &\quad + (1 - P_{1A})\bar{U}_1 + (1 - P_{2A})\bar{U}_2 \\ &\quad + \dots + (1 - P_{nA})\bar{U}_n \\ SEU_{\bar{A}} &= P_{1\bar{A}}U_1 + P_{2\bar{A}}U_2 + \dots + P_{n\bar{A}}U_n \\ &\quad + (1 - P_{1\bar{A}})\bar{U}_1 + (1 - P_{2\bar{A}})\bar{U}_2 \\ &\quad + \dots + (1 - P_{n\bar{A}})\bar{U}_n, \end{aligned}$$

where A and \bar{A} represent the two classes of actions, assertive or nonassertive. SEU is equal to the sum of the utilities of the possible positive and negative consequences weighted by their probabilities of occurrence, and the person should select the action with the larger SEU. The social component of the expectancy

model is treated as just another utility in the SEU model, which is what makes the latter a simpler model. Merging expectancy theory and the SEU model yields the prediction that the greater the SEU for an action, the more the person should intend to follow that course of action (BI). This model has been applied to a number of areas including career decision planning (Holmstrom & Beach, 1973; Mitchell & Beach, 1977; Muchensky & Fitch, 1975), third-grade children's decisions to attempt academic tasks (Gray, 1975), and family-planning decisions (Townes, Beach, Campbell, & Martin, 1977).

This study examines women's intent to act or not act in an assertive manner in light of their evaluations of the possible consequences of the two behaviors and the amount of risk perceived to be involved in performing either behavior given the characteristics of the situation. The strength of BI is measured by the number of times the participant states her intent to refuse a series of unreasonable requests that are made in videotaped or written scenes involving a high or low status male or female antagonist. Separate judgments of utility and probability permit computation of SEUs for assertion and nonassertion; persons differing in BI should have corresponding differences in their SEUs.

Method

General Strategy

First, participants were administered an assertiveness test and an anxiety test. Next they were presented with nine scenes (via videotape or written scripts) in which a male or female, authority or peer, made an unreasonable request. They were then given a list of 15 possible consequences of assertiveness/nonassertiveness and were asked to rate the utility (desirability) of each. Then the scenes were presented again, and after each scene the participants assessed their subjective probabilities that each of the aforementioned consequences would eventuate should they act either in an assertive or nonassertive manner. Finally, for each scene they were asked whether they would in fact comply or refuse.¹ The utilities and subjective probabilities were used to compute a SEU for each

¹ Clearly, it is also possible to choose not to act in a conflict situation (e.g., to wait for further information), but for convenience sake we have assumed two general and mutually exclusive classes of behavior even if this is an oversimplification.

scene for each participant, and the comply-refuse statements were used to compute an overall index of BI for each participant.

Participants

Sixty-four women attending undergraduate psychology classes and 47 women in a professional dental hygiene program participated in this study. The women were asked to attend one session lasting about 1½ hours and were seen in small groups. Following the experiment they were either given credit or were offered a short assertiveness training course.

Procedure

Participants were told that this study concerned decision making in difficult interpersonal situations; the term *assertiveness* was not mentioned until the debriefing explanation following the experiment. First they were given the Rathus Assertion Inventory (Rathus, 1973) and the Trait scale of the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970). They were given a list of 15 positive and negative consequences (feelings and actions) that might result from interpersonal conflict. This list was developed by presenting six assertiveness trainers with three sample situations and asking them to generate lists of positive and negative consequences commonly experienced by clients. The resulting core list of consequences was similar to those suggested by Alberti and Emmons (1970).

Participants reviewed the nine video or written scenes rapidly to give them an idea about what they were like and then rated their utilities for the consequences. They were asked first to decide if the consequence was positive or negative and then to show how positive by marking one end of a scale that ran from 1 to 4, or if negative, how negative by marking the other end of the scale from -4 to -1. A response of 0 indicated that the consequence was neutral.

They then were given a booklet of nine identical response sheets. The sheets were divided into two sections, the first for compliance and the second for refusal. Each was followed by a list of 10 consequences, which were generated from the core list used for the utility ratings. They were told that for each scene individually, they were to mark on a scale from 1 to 100 how probable each consequence would be if they personally were to comply and then how probable each would be if they personally refused to comply. (This method has been used successfully by Holmstrom & Beach, 1973; and Muchensky & Fitch, 1975.) When the probabilities were completed, participants were asked to mark whether they thought that they *actually* would comply or refuse the request; the proportion of refusals over the scenes was taken as the overall BI measure for each person.

The Scenes

Eight of the nine scenes that were used were from among those developed by Nedelman (1977) for a

study of the generalization of assertive behaviors; all eight scenes involved an unreasonable request. A ninth scene, involving a reasonable request, was included to see if there was a tendency for subjects to refuse to comply regardless of the appropriateness of the request; results for this scene revealed no such bias and responses to it are not included in the data analysis. The scenes and characters were described by a female narrator. Then there followed a direct request by the antagonist to the participant. Antagonists were either males or females, presented as either authorities or peers; each participant saw two examples of the four male-female/authority-peer combinations. For each session the sequences of scenes were presented in a scrambled order, with the restriction that each scene was presented as the first one in one of the sessions to spread the effect of presentation order.

Half of the participants saw videotape dramatizations of the scenes, and half received the same scenes as written scripts in booklet form.

Analysis of SEU

An SEU was computed for each scene by multiplying rated utilities by the rated probabilities of each of the 20 consequences. The SEUs were computed separately for compliance and for refusal and were then combined to yield a single score for each scene by subtracting the smaller SEU from the larger and assigning a minus sign to the difference if the larger SEU was for compliance and a plus sign if it was for refusal to comply. This was done so that there would be only one SEU datum for each scene rather than two. (Two would be awkward to analyze.) SEUs for each pair of similar scenes were averaged giving each participant four SEUs, one for each of the variations—male authority, female authority, male peer, and female peer. First, participants were divided into three equal-sized groups according to their scores on the assertion measure (high, medium, and low). An analysis of variance for repeated measures was performed to compare groups' SEUs and to examine the effects of the sex and status of the antagonist on SEUs. Then the analysis was repeated dividing the participants into three equal-sized groups according to their scores on the anxiety measure (high, medium, and low). Finally, participants were again divided into three equal-sized groups on the basis of BI (high, medium, and low), and the analysis was repeated. Significant interactions were further analyzed using Duncan's multiple-range test, and only significant effects will be discussed.

Results

Assertiveness and Anxiety

The first analyses were to see whether measured assertiveness and/or measured anxiety were related to BI and to each other. Table 1 contains the intercorrelations among these three variables for the psychology

students and the dental hygiene students separately. (Later analyses show the two groups to be significantly different.) Neither the assertiveness nor the anxiety test scores were significantly related to BI, and they were only slightly negatively related to each other.²

The SEUs

As was previously explained each participant's SEU was computed for each of the four different kinds of scenes. Because these SEUs are situation specific, it is inappropriate to combine them; thus there was no single overall SEU for each of the participants that could be correlated with their single BI measures. Therefore, participants were divided into psychology and dental hygiene students, and each group of students was divided into three equal-sized groups on the basis of the BI measure: high, medium, and low. Then a $3 \times 2 \times 2 \times 2 \times 2$ repeated measures analysis of variance was performed on the SEUs, with the independent variables of BI (high, medium, and low refusal to comply), kind of student (psychology or dental hygiene), method of presentation of the scenes (video or written), and the variables that defined the scenes; the status of the antagonist (authority/peer); and sex (male/female). Similar analyses were performed using the assertiveness test scores and the anxiety test scores instead of BI, both analyses showed these two variables to be unrelated to the SEUs, so they will not be discussed further.

Table 1
Multiple Correlations Among the Assertiveness Test Scores, the Anxiety Test Scores, and Behavioral Intent (BI) for Two Kinds of Students

Group	R
Psychology students	
Assert and anxiety	-.25*
Assert and BI	.18
Anxiety and BI	.03
Dental hygiene students	
Assert and anxiety	-.28*
Assert and BI	.26*
Anxiety and BI	-.17

* $p < .05$.

Table 2

Interaction Effects of Behavioral Intent (BI), Method of Presentation, Sex and Status of Antagonist Mean Subjective Expected Utilities

Variable	Status		Sex	
	Authority	Peer	Male	Female
BI				
High	11.44	10.54		
Medium	11.23	7.96		
Low	6.21	1.97		
Presentation				
Video			6.80	5.24
Written			10.61	7.57
Sex				
Male	11.15	6.83		
Female	7.28	5.87		

Note. For Method of Presentation \times Sex of Antagonist, $F(1, 94) = 4.1$, $p < .05$. For Sex \times Status of Antagonist, $F(1, 94) = 6.6$, $p < .01$. For BI \times Status of Antagonist, $F(2, 94) = 5.7$, $p < .01$.

For the BI analysis all main effects and three interactions (BI \times Status, Method of Presentation \times Sex of the Antagonist, and Status \times Sex of the Antagonist) were significant.

The first main effect is BI, $F(2, 94) = 9.6$, $p < .01$, with a mean SEU for the high BI group of 10.99, for the medium BI group of 9.60, and for the low BI group of 4.09. Using Duncan's test the difference between the means of the high and low BI group was significant at $p < .05$. The medium BI group did not differ from either the low or high BI groups.

For the second main effect, kind of student, $F(1, 94) = 6.8$, $p < .05$, the psychology students has a group mean SEU of 6.67 and the dental hygienists' mean was 9.42. This shows that the SEUs for the latter favored refusal more than did the SEUs for the psychology students. Further, the mean BI for the hygienists (1.72) was significantly

² In light of this result for the Rathus scale, readers familiar with this scale may be interested in the obtained means and standard deviations (for psychology undergraduates, $M = 12.62$, $SD = 20.49$; for dental hygiene students, $M = 10.60$, $SD = 22.41$).

higher, $t(104) = 2.80$, $p < .01$, than the mean BI for the psychology students (1.44), indicating that the hygienists said that they would refuse to comply with the requests in the scenes more often than the psychology students did.

The third main effect, method of presentation of the scene, $F(1, 94) = 10.0$, $p < .01$, had a SEU for the written presentation (9.09) that was higher than for video presentation (6.02). The corresponding mean BIs were 1.55 for the written presentation and 1.50 for video presentation, $t(104) = .5$, *ns*. These results suggest that BI is independent of the method of presentation but that the participants' expectations for refusal (SEU) were more positive for those who saw the written scripts, and that those who saw the videotapes—a more realistic stimulus—were more likely to have less favorable expectations for refusal.

The fourth and fifth main effects were for the status of the antagonist, $F(1, 94) = 27.3$, $p < .01$, with mean SEUs of 9.22 for authorities and 6.35 for peers and for the sex of the antagonist, $F(1, 94) = 45.6$, $p < .01$, with mean SEUs of 8.99 for males and 6.58 for females. Participants' expectations for assertion to peers and to women were more negative than for assertion to authorities and males. Because these two variables define the scenes rather than divide participants into groups, as did the student and presentation variables, BIs could not be calculated for each of the four variants for comparison with the SEUs. However, it is possible to calculate the proportion of times that the participants stated an intent to refuse to comply for each level of the two variables. For authorities the proportion of refusal was .85, for peers it was .69, for males it was .77, and for females it was .76. Because these proportions each involve multiple contributions by all of the participants, no statistical tests can be performed. However, the proportions appear to be congruent with the SEU for status and less so for sex.

All three significant two-way interactions involved the status and sex variables. These are shown in Table 2.

Within the different kinds of scenes, the participants in the high and the low BI

groups differed significantly in their SEUs for refusal to comply in scenes, although this is not true of their response to the scenes involving authority figures. For both scene types, as one moves from high to low BI, the SEUs for refusal to comply to authority and peer demands become increasingly disparate, though these differences did not reach statistical significance. For those persons who have a high BI to refuse, the difference between refusal to peer and to authority was only .90; for the medium group it was 3.27, and for the low group it was 4.24. For the participants with a low BI to refuse, unlike the high group, it appears much more difficult to refuse requests made by peers than by authorities.

The interaction between method of presentation and sex of antagonist shows that the SEUs for written presentation were higher overall than the SEUs for the video presentation for both scene types. SEUs for refusal to comply with males were higher than SEUs for refusal to comply with females for both methods of presentation.

The interaction between sex and status of the antagonists shows that the SEUs for refusal were higher for male authorities than for female authorities, male peers, and female peers. The SEUs for female authorities were higher than for female peers, with no differences between the remaining categories. The corresponding proportions of refusal to comply statements (BI) for the four kinds of antagonists are .88 for male authorities, .82 for female authorities, .67 for male peers, and .71 for female peers—moderate agreement with the SEUs. In sum, it appears that the highest SEUs for refusal occurred among the dental hygiene students, for participants with high BI, and when participants are confronted with authority figures, particularly males.

Utility and Probability

Because SEU is clearly related to BI, it is appropriate to see if the differences between participants who frequently state that they intend to refuse to comply and those who less frequently state intention to refuse can be attributed to differences in either their utilities for the possible outcomes or their subjective probabilities that these outcomes

will occur. To do this the participants were divided into groups on the basis of the kind of student (psychology, dental hygiene), method of presentation (video, written), and BI (high, low). A repeated measures analysis of variance using utilities as the dependent variable yield no significant effects of any of the independent variables. As can be seen in Figure 1, students at both levels of BI evaluated the utilities in the same way.

A similar analysis was performed using the probabilities as the dependent variable. Here all three variables (type of student, method of presentation, and BI) had a significant effect on the probability estimates, and all interactions also were significant. This is illustrated for high and low BI participants in Figure 2.

The graph shows that if they were to comply, the low BI participants would expect the positive consequences to be more probable, $t(62) = 6.08, p < .01$ than would the high BI participants. Similarly, if they were to refuse to comply, the low BI participants would see the positive consequences as less probable, $t(62) = 8.33, p < .01$, and the negative consequences as more probable, $t(62) = 4.29, p < .01$, than the high BI participants. In short, the difference between high and low BI participants is not in their utilities for the positive and negative consequences but in their different perceptions of the probabilities of those positive and negative consequences occurring should they elect to comply or to refuse to comply.

Discussion

Schools and Presentations

Initial analyses revealed significant effects for the type of students and for the method of presentation of material. SEUs for the dental hygiene group were uniformly higher than SEUs for the psychology students. Attitudes of the former toward the requests in the scenes as suggested by their higher intent to refuse may have been mediated by their expectation that they would soon be employed in a highly authority-oriented profession. The experience and advice of those few who had already worked as dental assistants contributed to this concern.

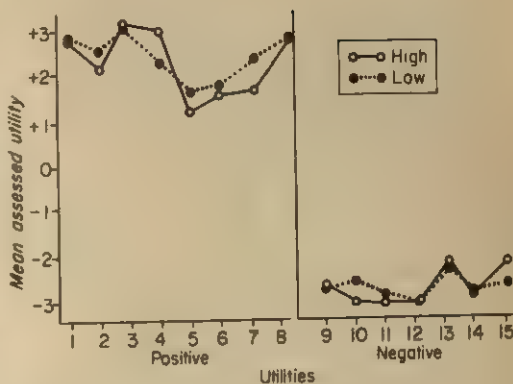


Figure 1. Utilities for participants who were high and low in behavioral intent.

The effect of presentation was also significant. SEUs for the written scenes were higher for both sex and status and across schools, although the pattern of refusal to scenes did not differ. Even though the written scenes allowed participants to make greater use of their own experience in evaluating the scenes, the demand of the antagonist was less immediate. Participants evaluated the consequences for refusal as more negative when the request was made via videotape. Videotape has been a successful adjunct to treatment in assertive training studies for modeling appropriate responses (Hersen, Eisler, Alford, & Pinkston, 1973), and for assessing change (Hersen, Eisler, & Miller, 1973). Although it has not been systematically tested in the literature, it is likely to be a better medium than audiotape or written instructions, since the stimulus it presents apparently is experienced as "more real."

The Scenes

Although many studies acknowledge the variation of task difficulty involved in assertion, no systematic investigation of the subject's evaluation of the risks involved exists in the literature. Typically authors deal with this issue by hierarchically ordering materials differing in risk or task difficulty (Alberti & Emmons, 1970; Piaget & Lazarus, 1969; Wolpe & Lazarus, 1966). Situations involving interactions with persons of the same sex, peers, or persons who have a distant relationship are presented prior to those involving persons

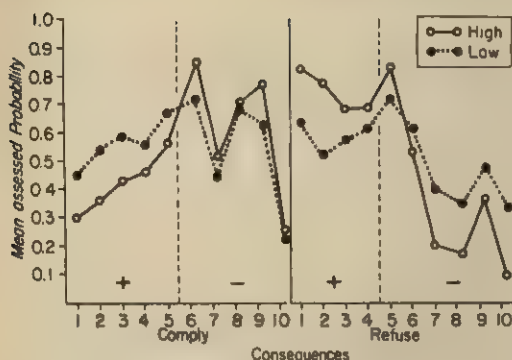


Figure 2. Probabilities for participants who were high and low in behavioral intent.

of the opposite sex, spouse, or an authority figure (McFall & Lillisand, 1971). The rationale for this is to provide success in situations involving less risk of failure and emotional concomitants and to desensitize the anxious patient to difficult interactions. This pattern has become institutionalized, but, as was demonstrated in this study, it may not reflect the order of task difficulty for all participants. Perhaps since all subjects were students and were less likely to be confronted with authority figures in the work place, it is possible that they feared the consequences of assertion to authorities less than they feared "negative" consequences in relationships with peers. If this is the case, a training program designed to teach assertion should take into account the participants' own evaluation of the salient perceived risks involved in the new behavior and the current situation. It is likely that these may differ from participant to participant and from established training procedures.

Behavioral Intent

The measure of intent to refuse (BI) was highly related to SEU for the content scenes. This attests to the situational specificity of assertion. It suggests that assertiveness is not a trait that applies equally across situations varying in sex and status of the antagonist and presumably varying also in the amount of risk entailed.

In this study BI is not to be taken as an actual indication of the participant's willingness to assert in a real-life situation. However,

a number of studies have found a high correlation between behavioral intent and actual behavior in a variety of situations (reviewed in Fishbein & Ajzen, 1975). These authors assert that intent is the best indication of actual behavior, but they qualify its ability to predict behaviors in situations in which there is a time interval between the intent and the behavior, and when the behavior in question depends on the actions of others. Additionally, when the stimuli are highly artificial, as in this experiment, intent may be no more than an indication of the subjects' desire to behave in the intended manner.

The fact that there were no systematic differences in SEU for participants scoring high and low on the personality measures may have been a function of the tests used, and they may also reflect the inadequacy of assigning a "trait" value to a situation-specific task. A major issue in the literature has been the specificity of the assertiveness response and the expectations that can be made for generalization. Hersen, Eisler, and Miller (1973) and Wolpe and Lazarus (1966) argued that assertion is specific to the social context in which it occurs. Others have found some degree of transfer of training as a result of their intervention (McFall & Lillisand, 1971; McFall & Marston, 1970; Nedelman, 1977).

Utility and Probability Components

Participants high and low in BI to refuse unreasonable requests did not differ in the way they valued the utilities of the consequences. However, participants differed significantly in their perception of the probabilities of the consequences occurring as a result of their behavior. For refusal, these trends were reversed; participants with high BI for refusal saw the positive consequences of refusal as more likely to occur and the negative consequences as less probable. For both events the probability endorsements of low BI participants were generally more constricted, whereas high BI subjects gave more extreme probabilities.

This result is consistent with results reported by Mitchell and Knudsen (1973), who separated out the effects of utilities and probabilities.

ities. Comparing college students and business students, they found that although the students did not differ in their valuation of occupational goals, they held significantly different probabilities that a career in business would allow them to attain those goals.

Findings support the hypotheses that intent to refuse to comply with an unreasonable request depends on the attributes of the situation. Additionally, subjects did not differ in their valuation of the positive and negative consequences, but instead they differed in their expectations that a particular consequence would occur should they elect to refuse to comply with unreasonable demands.

Issues for Training

Research on assertiveness has had its greatest impact on applied methods for clients. Past studies (Hersen, Eisler, & Miller, 1973; McFall & Marston, 1970) have identified behavioral components of assertiveness that can be taught to nonassertive persons and used in a variety of situations. In the present study it appears that participants scoring on all levels of an assertion inventory also respond to situations differentially according to the sex and status of the antagonist and presumably according to the degree of risk that they perceive to be a consequence of their behavior.

In designing a comprehensive assertiveness training program, these results suggest that the focus of training should be on changing the participant's cognitive expectations about the results of his or her behavior, as well as on changing attitudes or specific behaviors. This approach assumes that the average client has greater knowledge about appropriate interpersonal behaviors than might be found among the hospitalized schizophrenic male patients in the Hersen et al. studies. In fact, the thrust of training has not been directed toward this latter group but rather toward women and college students who may have adequate skills but who do not act assertively when they perhaps should. The Gottman and Schwartz study (Note 1) suggests that both assertive and nonassertive people know the appropriate behaviors. It may be that the primary "skill deficit" lacking in nonassertive

persons is their inability to accurately estimate the consequences of their assertion. If this is true, greater emphasis must be placed on changing the person's cognitive expectation of the consequences. This may involve a process of teaching the client a new set of expectancies about possible outcomes based on the characteristics of differing situations. Since information affects beliefs about consequences, active participation such as role playing is likely to be an effective strategy for change, because it exposes the person to new information and allows for change in belief. Additionally, as suggested by Mitchell and Biglan (1971), exploring the components of the client's expectancies in therapy might change the client's perception of instrumental relations between behavior and outcomes or the evaluations of the probabilities of outcomes occurring.

This research also suggests that training programs might profit from a broader appreciation of the client's negative as well as positive evaluations of the consequences of assertion. The weighting of these consequences appears to differ from situation to situation and may determine the persons' willingness to respond. Assertive behavior may be realistically evaluated by the subject as counterindicated in a given situation. For example, in some situations compliance to a mildly unreasonable request may have the secondary positive consequences of strengthening a friendship. In other situations the effect might be to advance a long-range goal. It may be necessary, as Hersen, Eisler, and Miller (1973) suggested, to train for generalization of assertive behavior, in this case to teach clients appropriate expectations in a variety of situations. It may also be important to teach relevant others to change their expectations as the client changes his or hers. Since we have only treated one type of assertion in this study, refusal to comply with an unreasonable request, further investigation is needed to see whether other assertive responses (complimenting, standing up for one's opinions, etc.) are similarly affected by the situational variables and by the participant's expectations of the consequences.

A number of issues have been raised in this study that have not been dealt with in the

literature of assertion and that seem worthy of further investigation. These include an appreciation of the client's perception of the risk involved. The traditional pattern of training female peer situations first because they involve less risk than authority situations does not take into account the possible ascendance of peer and friendship values in a female college population. In an analogue study of this type, the use of self-report and self-ratings are limited in their ability to predict actual behavior. It appears, however, that the decision/expectancy model has value as a research tool in specifying under what conditions a participant might choose to act in an assertive manner in a given situation. Future research might investigate the ability of this approach to predict and shape the responses of individuals in real-life situations.

Reference Note

1. Gottman, J., & Schwartz, R. *Toward a task analysis of assertive behavior*. Unpublished manuscript, Indiana University, 1976.

References

- Alberti, R. E., & Emmons, M. L. *Your perfect right—a guide to assertive behavior*. San Luis Obispo, Calif.: Impact, 1970.
- Edwards, W. The theory of decision making. *Psychological Bulletin*, 1954, 51, 380-417.
- Fishbein, M., & Ajzen, I. *Belief, attitude, intention and behavior*. Reading, Mass.: Addison-Wesley, 1975.
- Gray, C. A. Factors in students' decisions to attempt academic tasks. *Organizational Behavior and Human Performance*, 1975, 13, 147-164.
- Hersen, M., Eisler, R., Alford, H., & Pinkston, S. Effects of practice, instructions and modeling on components of assertive behavior. *Behaviour Research and Therapy*, 1973, 11, 443-451.
- Hersen, M., Eisler, R., & Miller, P. Development of assertive responses: Clinical measurement and research considerations. *Behaviour Research and Therapy*, 1973, 11, 505-521.
- Holmstrom, V., & Beach, L. R. Subjective expected utility and career preferences. *Organizational Behavior and Human Performance*, 1973, 10, 201-207.
- Lazarus, A. A. *Behavior therapy and beyond*. New York: McGraw-Hill, 1971.
- McFall, R. M., & Lillisand, D. B. Behavior rehearsal with modeling and coaching in assertive training. *Journal of Abnormal Psychology*, 1971, 77, 313-323.
- McFall, R. M., & Marston, A. R. An experimental investigation of behavior rehearsal in assertive training. *Journal of Abnormal Psychology*, 1970, 76, 295-303.
- Mitchell, T. R., & Beach, L. R. Expectancy theory, decision theory, and occupational preference and choice. In M. F. Kaplan & S. Schwartz (Eds.), *Coordination and control of group and organizational performance*. New York: Academic Press, 1977.
- Mitchell, T. R., & Biglan, A. Instrumentality theories: Current uses in psychology. *Psychological Bulletin*, 1971, 76, 432-454.
- Mitchell, T. R., & Knudsen, B. Instrumentality theory predictions of students' attitudes toward business and their choice of business as an occupation. *Academy of Measurement*, 1973, 16, 41-52.
- Muchensky, P. M., & Fitch, M. K. Subjective expected ability and academic preferences. *Organizational Behavior and Human Performance*, 1975, 14, 217-226.
- Nedelman, D. *An experimental analysis of the dimensions of generalization of assertive training*. (Doctoral dissertation, University of Washington, 1976). *Dissertation Abstracts International*, 1977, 37, 1444B. (University Microfilms No. 76-20, 728).
- Piaget, G. W., & Lazarus, A. A. The use of rehearsal-desensitization. *Psychotherapy: Theory, Research, and Practice*, 1969, 6, 264-266.
- Rathus, S. A 30-item schedule for assessing assertive behavior. *Behavior Therapy*, 1973, 4, 398-406.
- Spielberger, C., Gorsuch, R., & Lushene, R. *Manual for the State-Trait Anxiety Scale*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Townes, B. D., Beach, L. R., Campbell, F. L., & Martin, D. C. Birth planning values and decisions: The prediction of fertility. *Journal of Applied Social Psychology*, 1977, 7, 73-88.
- Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford, Calif.: Stanford University Press, 1958.
- Wolpe, J. *The practice of behavior therapy*. New York: Pergamon Press, 1969.
- Wolpe, J., & Lazarus, A. A. *Behavior therapy techniques: A guide to the treatment of neuroses*. New York: Pergamon Press, 1966.

Received August 10, 1977 ■

Use of Paradoxical Intention in a Behavioral Program for Sleep Onset Insomnia

L. Michael Ascher and Jay S. Efran
Temple University

Sleep onset insomnia seems often to be based on performance anxiety associated with a client's fears of being able to fall asleep; in some cases, a therapeutic program might actually exacerbate this performance anxiety by focusing on the client's efforts to voluntarily control the sleep onset process. Five cases of sleep onset difficulty, unusually resistant to a conventional behavioral program for this problem (i.e., deep muscle relaxation and systematic desensitization), were exposed to paradoxical intention suggestions requiring that they try to remain awake as long as possible, rather than attempt to fall asleep. A rapid reduction of sleep onset latency occurred following the shift from the conventional program to the paradoxical intention instructions.

Frankl (1975) has recently provided a more detailed and updated description of "paradoxical intention," a psychotherapy technique that he developed in the framework of logotherapy. He has also begun to articulate some of the links between this technique and behavior therapy approaches. This article reports the use of paradoxical intention as a complement to other behavioral procedures in cases of sleep onset "insomnia" that proved resistant to treatment.

Deep muscle relaxation as developed by Jacobson (1938) and modified by Wolpe (1958, 1974) has been used with success in sleep disturbances due to other than physiological factors (Kahn, Baker, & Weiss, 1968; Nicassio, & Bootzin, 1974). There are, however, a significant number of clients with whom relaxation alone is not sufficient to produce the desired degree of improvement. In such cases, behavior therapists have commonly used systematic desensitization focused either on the anxiety-provoking themes that occupy the client prior to sleep onset or toward anxiety directly related to the sleep situation (Borkovec, Steinmark, & Niu, 1973; Geer & Katkin, 1966). Other behavioral methods, such

as covert conditioning, thought stopping, and operant stimulus control have also been used singly and in combination.

The present article explores the use of paradoxical intention as an ancillary treatment with individuals for whom the relaxation-desensitization program seemed insufficient. It is relatively easy to apply and usually produces immediate behavior changes without the complications of medication (Kales, Allen, Schaff, & Kales, 1970). Briefly, paradoxical intention can be viewed as a behavioral prescription requiring clients to perform responses that appear to be incompatible with the goal for which they are seeking therapeutic assistance. Thus, in the present context, clients with sleep onset disturbance are requested to try to *remain awake* for as long as possible, rather than to focus on trying to fall asleep. In other words, they are asked to exaggerate the very behavior that they would like to reduce. In the present study, the course of treatment of five clients illustrates the use of paradoxical intention within a behavioral framework when the sleep difficulty did not yield to a conventional relaxation-desensitization program.

Method

Subjects

All five clients had applied for treatment to the clinic of the Behavior Therapy Unit of the Department of

Requests for reprints should be sent to L. Michael Ascher, Behavior Therapy Unit, Eastern Pennsylvania Psychiatric Institute, 3300 Henry Avenue, Philadelphia Pennsylvania 19129.

Table 1
Characteristics of the Five Clients

Client	Sex	Age	Marital status	Occupation	Average sleep onset latency (minutes)	Length of complaint (years)	Primary presenting problem
A	Male	32	Single	Lawyer	60	5	Primary erectile dysfunction
B	Male	27	Married	Graduate student (mathematics)	45	12	Interpersonal difficulties
C	Female	41	Married	Housewife	90	21	Sleep onset
D	Female	23	Married	Social worker	45	3	Sleep onset
E	Male	25	Single	Salesman	75	4	Sleep onset

Psychiatry at Temple University. The first author served as the therapist and conducted weekly sessions individually with each client. The clients are further described in Table 1.

Procedure

Two weeks prior to initiating therapy directed at sleep onset difficulty, each client was asked to chart the approximate length of sleep onset latency each morning in addition to recording other details relevant to the sleep situation (e.g., mood when retiring, time of retiring, number of awakenings, "restfulness" of sleep).¹ Beginning with the initial session following this 2-week baseline period, and continuing for 10 weeks, all clients were instructed in deep muscle relaxation (Wolpe, 1974) with particular emphasis on the use of these exercises prior to retiring. Clients were also advised on the modification of their sleeping arrangements to produce optimal conditions for rapid sleep onset. With successive sessions, additional techniques were introduced as required (e.g., desensitization and covert conditioning). (Although the behavioral program outlined above has been of demonstrated efficacy with sleep onset difficulty, it failed to produce the desired improvement in the five cases reported in the present study. This represents about 10%-15% of individuals seen for sleep disturbance problems over a 4-year period.)

During the first session following the 10-week period, the therapist suggested that a modification in the procedure might enhance the patient's progress. Paradoxical intention was administered by instructing the client (with an appropriate rationale) *to try to remain awake*. Three clients (A, B, and E) were told that although the present program would eventually prove to be successful for them, specific details would be required to "further tailor therapy to meet their individual needs." These clients were told that the failure of the program at this point was possibly due to the lack of sufficient information regarding the sleep situation. It was suggested that since detailed descriptions of thoughts experienced just before sleep were required, they should try to remain awake as long as possible to experience these thoughts. If they fell asleep too soon, the necessary information would be unobtainable.

With the remaining two clients (C and D), it was

suggested that the relaxation component of the behavioral program was not of sufficient duration to "produce the level of relaxation requisite for sleep." Therefore, instructions were given to lengthen the number of steps (and, consequently, the length of time) required to complete the relaxation practice. Clients were advised to go through the entire procedure several times to achieve a satisfactory level of relaxation. They were asked to do this even if it meant resisting the urge to sleep.

The paradoxical intention procedure was continued for all five clients for 2 weeks. During the sessions the clients were asked how they had progressed with respect to the assigned tasks (either obtaining the thoughts experienced prior to sleep or increasing the relaxation procedure to achieve a "deeper" level of relaxation). Typically, they reported that they had not been able to accomplish the goal because they had fallen asleep too quickly. The therapist briefly expressed interest and encouragement on hearing this information but suggested that they redouble efforts to accomplish the goal, trying harder to remain awake. This entire interaction took only a short time during the initial portion of the three relevant therapy sessions, the remainder of which were devoted to continuation of the regular desensitization procedures.

Following the 2-week paradoxical intention period, four of the clients (A-D) were given no further sleep onset treatment. However, Client E was instructed to return to his previous program, which incorporated techniques focused on the reduction of anxiety-provoking thoughts experienced during the sleep onset latency period (i.e., thought stopping, covert positive reinforcement, systematic desensitization). The client remained on this program for 3 additional weeks, after which he was told that his efforts to reduce discomforting thoughts seemed to have been effective and that he should reemphasize the relaxation (paradoxical) component of the program. In this way, paradoxical intention instructions were again administered (this time

¹ These other measures seemed less relevant in this context than the latency measure, and, for the sake of brevity, they have been omitted. All of the measures followed the same pattern.

Table 2
Mean Number of Minutes to Sleep Onset for Each Week of the Treatment Program

Mean Number of Minutes to Sleep Onset for Each Patient																				
Week	Phase 1			Phase 2						Phase 3			Phase 4			Phase 5				
	1	2	M	1	2	3	4	5	6	7	8	9	10	M	1	2	M	1	2	M
	Patient																			
A	36	40	38	34	28	31	26	31	30	27	33	25	35	30	14	10	12	17	12	14.5
B	32	25	28.5	31	24	26	33	29	27	33	23	34	23	28.3	4	7	5.5			
C	84	96	90	63	71	68	62	80	74	79	67	64	69	69.7						
D	26	32	29	29	25	22	28	31	33	26	30	32	29	28.5	11	15	13			
E	59	56	57.5	36	47	42	39	43	38	40	40	37	41	40.3	7	5	6			
M	47.4	49.8		38.6	39	37.8	37.6	42.8	40.4	41	38.6	38.4	39.4		10.6	9.8				
SD	21.4	25.3		12.4	18.1	16.5	13	19.3	17.2	19.7	15.2	13.4	15.9		4.7	3.5				

Note. Phase 1 = baseline; Phase 2 = conventional program; Phase 3 = paradoxical intention; Phase 4 = readministration of conventional program; Phase 5 = readministration of paradoxical intention.

Results

Table 2 presents the mean self-report estimations of sleep onset latency (i.e., the duration between "lights out" and sleep onset) for each client during the 2-week baseline period (Stage 1), the 10-week behavioral program administration (Stage 2), and the 3-week paradoxical intention period (Stage 3), which represented the terminal stage for four clients (A-D). In addition, data are reported for Client E's second behavioral administration (Stage 4) and return to paradoxical intention (Stage 5). A comparison of the sleep onset latencies during the baseline period with the data following 10 weeks of behavior modification indicates that in most cases "some improvement" was obtained, even though it was judged insufficient by the client.³

The data show a marked reduction in sleep onset latency following the administration of paradoxical intention instructions to "try to remain awake." In the case of Client E, the data for the first three stages of the study are congruent with those for the remaining clients. Sleep onset latency was somewhat reduced as a result of relaxation training and behavioral procedures directed at distracting anxiety-provoking thoughts. However, presentation of paradoxical intention instructions produced a marked reduction in sleep onset latency. Reinstitution of the previous behavioral program was coincident with an increase in sleep onset latency, which decreased again when paradoxical intention instructions were readministered. Informal long-term follow-up (by telephone) indicated that each of the clients remained satisfied with their sleep behavior after a period of 1 year.

² The clients appeared to believe the treatment rationales that they were offered, although formal data on this issue could obviously not be collected in this clinical setting.

³ Formal statistical treatment is perhaps unwarranted in this study. However, the reader may wish to note that *t* tests between the baseline and behavior therapy, the baseline and paradoxical intention, and behavior therapy and paradoxical intention were, respectively, .978 (*ns*), 3.86 ($p < .02$), and 4.58 ($p < .02$).

Discussion

The present study illustrates the utility of paradoxical intention within the context of the behavioral treatment of sleep onset difficulties. A reasonable question would seem to be, why did paradoxical intention produce a change that the conventional treatment alone could not? Paradoxical intention has been shown to be effective with a wide variety of psychosomatic dysfunctions, that is, physiological processes, having autonomic nervous system innervation, which can be inhibited by anxiety. Such dysfunctions can occur, for example, with various aspects of sexual activity, elimination, and, as in the present case, components of sleep behavior.

Most people experience occasional difficulty in falling asleep. This difficulty is usually seen by the individual as an isolated event precipitated by unusual excitement or tension during the day, too much sleep prior to bedtime, and so forth. Succeeding evenings normally result in a rapid return of the individual's typical sleep pattern. However, a small percentage of people view instances of sleep onset difficulty as indicants of a trend toward decreasing levels of satisfactory functioning. This latter group considers each successive evening a test of their ability to fall asleep. The level of performance anxiety increases as each test approaches. Anxiety is assumed to have a reciprocally inhibiting relationship with sleep onset (as with sexual arousal); that is, it stimulates the sympathetic nervous system, a system that alerts the organism and is the reciprocal of the parasympathetic system, which has recuperative functions compatible with sleep onset.

Some people who experience performance anxiety at bedtime appear to focus on at least two major themes: First, they monitor their level of sleep readiness. Second, they focus on negative outcomes associated with sleep loss (e.g., difficulties that they may experience on the following day, physical well-being, etc.). For these individuals, concern about successful performance and contingencies of failure serve to increase anxiety prior to bedtime and main-

tain a high level of anxiety in the sleep situation. "Trying hard" to get to sleep only makes this cycle more pernicious.

The cycle can apparently be broken by paradoxical intention, which decreases performance anxiety by redefining the situation. The paradoxical suggestion is incompatible with the "common sense" effort that the individual has been making to perform the target response. Thus, because the conventional program was aimed at helping these individuals fall asleep more rapidly, it inadvertently supports the cycle of "performance anxiety—failure to perform—increased performance anxiety." Paradoxical intention removes the client from this system. In this article, paradoxical intention was used with difficult cases. It remains to be seen whether it would be as useful with the wider range of clients complaining of sleep onset latency.

References

- Borkovec, T. D., Steinmark, S. W., & Niu, S. D. Relaxation training and single item desensitization on the group treatment of insomnia. *Journal of Behavior Therapy and Experimental Psychiatry*, 1973, 4, 401-403.
- Frankl, V. E. Paradoxical intention and dereflection. *Psychotherapy: Theory, Research and Practice*, 1975, 12, 226-236.
- Geer, J. H., & Katkin, E. S. Treatment of insomnia using a variant of systematic desensitization: A case report. *Journal of Abnormal Psychology*, 1966, 71, 161-164.
- Jacobson, E. *Progressive relaxation*. Chicago: University of Chicago Press, 1938.
- Kahn, M., Baker, B. L., & Weiss, J. M. The treatment of insomnia by relaxation training. *Journal of Abnormal Psychology*, 1968, 73, 556-558.
- Kales, A., Allen, C., Schaff, M. B., & Kales, J. D. Hypnotic drugs & their effectiveness. All-night EEG studies of insomniac subjects. *Archives of General Psychiatry*, 1970, 23, 226-232.
- Nicassio, P., & Bootzin, R. A comparison of progressive relaxation and autogenic training as treatments for insomnia. *Journal of Abnormal Psychology*, 1974, 83, 253-260.
- Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford, Calif.: Stanford University Press, 1958.
- Wolpe, J. *The practice of behavior therapy*. New York: Pergamon Press, 1974.

Received September 26, 1977 ■

A Scoring Manual for Literalness in Proverb Interpretation

Chris A. Hertler, Loren J. Chapman, and Jean P. Chapman
University of Wisconsin—Madison

A scoring system is offered for literalness of proverb interpretation as an alternative to scoring concreteness. For a group of 115 schizophrenic and normal subjects, literalness and Gorham's concreteness were equivalent on coefficient alpha (.85 for literalness and .84 for concreteness). Interrater reliability was .90 for both scoring systems. Nevertheless, abstraction correlated lower ($p < .01$) with literalness than with concreteness. For 77 schizophrenics, Verbal IQ correlated significantly with concreteness ($r = -.52$, $p < .01$) but not with literalness ($r = -.15$, *ns*). Thus, literalness is less affected by intelligence and by ability to respond abstractly than is Gorham's concreteness.

When schizophrenics are asked to interpret proverbs, they often respond to the proverb as a literal statement rather than as a bearer of a figurative meaning. For example, when asked to interpret the statement "When the cat's away, the mice will play," even educated and intelligent schizophrenic patients may explain the actions of cats and of mice, rather than of people. This article offers a system for scoring literalness of proverb interpretation.

Dozens of writers have discussed literalness of proverb interpretation, but scoring systems treat it as a manifestation of concreteness. The scoring systems score either concreteness or abstraction, which is merely a term for accuracy of proverb interpretation that is incompatible with literalness. Both low abstraction and high concreteness are usually viewed as implying literalness. However, scoring for either concreteness or abstraction classifies literalness with various other kinds of poor performance that have little to do with literalness. Literalness is a more specific error than either

concreteness or lack of abstraction. Literalness is an active attempt to interpret the words of the proverb as a literal message rather than as symbols to be interpreted.

The most commonly used scoring system for proverbs is that of Gorham (1956a, 1956b), who scores abstraction and concreteness separately. He offered a detailed scoring system for abstraction but evidently regarded concrete responses as so obvious that no formal scoring system need be offered. Gorham (1956a) stated his criteria for concreteness very briefly.

Concrete answers are usually apparent to a clinical observer. They stick closely to the symbols of the proverb. In schizophrenics, it is common for patients to substitute "That's right," "exactly," "that is not so because," or "yes" and "no" for a restatement of the proverb in concrete form. These answers are considered to be concrete." (p. 3)

Gorham supplemented this statement with one example of a concrete response to each of seven proverbs.

Many responses by both normal subjects with low intelligence and schizophrenics stick closely to the symbols of the proverb but yet are not literal interpretations of the proverb. If a subject is unable to interpret the proverb but is verbose, he or she will talk about the symbols. Subjects who cannot interpret a proverb appropriately often simply repeat some of the words of the proverb without further elaboration, give associate responses to the words,

This article was supported by Research Grant MH-18354 from the National Institute of Mental Health, U.S. Public Health Service.

The authors are indebted to Richard Lang and to Robert Adamski for assistance with the scoring.

Inquiries concerning this article should be sent to Loren J. Chapman, Department of Psychology, W. J. Brogden Psychology Building, University of Wisconsin, 1202 West Johnson Street, Madison, Wisconsin 53706.

relate the words to their own experience, or talk in other discursive ways about the proverb that they are asked to interpret. Such talk is not evidence that he or she interprets the symbols literally, but it would be scored as "concrete" by Gorham's criteria. For example, one schizophrenic responded to "The worst spoke in the wheel breaks first" with "Wheel breaks, brake locks, break off." Another patient responded to "He who stumbles twice over one stone deserves to break his shins" with "I don't stumble, walk straight, never stumble, got to control me, can't see." Both of these responses stick closely to the words of the proverb and, hence, would be scored as concrete, but they are not attempts at a statement of a literal meaning. Gorham's concreteness seems to reflect in large part dullness and a lack of accuracy. In schizophrenia, concreteness is heavily affected by a failure to focus on the task of interpretation and by other aspects of generalized deficit. Literalness should be less affected by generalized deficit because it is a more specific kind of error. Because literalness is less a reflection of generalized deficit than is concreteness, a score for literalness should depend less than does concreteness on both Verbal IQ and abstraction.

The system of scoring literalness that we offer here labels all appropriate answers as nonliteral and only some incorrect answers as literal. This system provides for scoring each proverb on a 3-point scale. This follows from dividing the proverb into two halves, each of which could receive a literalness score of 0 or 1. Thus the total literalness score for the sum of the two halves could be 0, 1, or 2. For example, the proverb "Rome was not built in a day" is sometimes interpreted as "Rome took a long time to build." In this response, *Rome* is treated literally, and *in a day* is treated abstractly, which yields a total literalness score of 1.

The scoring principles offered here could be applied to the interpretation of any figurative statement, although we developed them for proverbs from Gorham's (1956a, 1956b) clinical form of the Proverbs Test. Actually, not all 36 of Gorham's items are proverbs in the sense of being figurative statements to be interpreted. Some items are, instead, aphorisms that should be interpreted literally. Examples

are "The more cost, the more honor" and "Where there's a will, there's a way." We do not include such items in our scoring system. We regard 24 of Gorham's items as clearly figurative statements and, therefore, include them. These are Form I, Items 2-9 and Item 12; Form II, Items 1-7 and Item 10; Form III, Items 1, 2, 5, 6, 7, 9, and 11.

Following the example of Friedes, Grisell, Levin, Dobie, and Cohen (Note 1), we designate certain of the words in each half of the proverb as symbols that must be generalized or interpreted to obtain a correct abstract interpretation. For example, in the proverb, "A drowning man will clutch at a straw," *drowning* and *straw* are symbols that must be interpreted, but *clutch* is not. If *drowning* or *straw* are repeated in the answer, a score for literalness must be considered. However, the appearance of *man* or *clutch* in the answer need not imply literalness. For example, the response "A man who is in trouble will clutch at any method to save himself" is an adequate abstract interpretation rather than a literal one.

Scoring Principles for Literalness

For the sake of brevity, we will illustrate each scoring principle with responses to the proverb "Rome was not built in a day." The two halves of this proverb are *Rome* and *was not built in a day*. The symbols to be generalized in an abstract response are *Rome*, *built*, and *day*.

An entire proverb is considered completely unscorable for literalness if the entire response consists of any of the following:

1. An "I don't know," without further elaboration.
2. A reference to a personal experience of the subject as a substitute for interpreting the proverb, for example, "I have never been to Rome."
3. A response that has no recognizable relationship either to the literal meaning of the proverb or to a possible interpretation of the proverb. Responses can be judged as falling in this category even if they contain one or more of the symbols of the proverb, for example, "Rome is in Italy."
4. A repetition of the proverb without further elaboration, for example, "Rome was not built in a day."

5. A repetition of only part of the proverb without further elaboration, for example, "Built in a day."

6. A semantic associate or a clang associate to one of the symbols without further elaboration, for example, "Paris," or "Cathedral domes."

7. Any single word other than yes or no and other than an equivalent to yes or no such as absolutely. An example of the unscorable response is *Italy*.

8. A bizarre or autistic response, with or without further elaboration, for example, "Roman vices accentuate carnal lust" or "Roman vices can't be learned quickly."

9. No response whatever.

Note, however, that many of these kinds of responses are scored if the subject adds other words in the response. See examples below.

A proverb receives a total literalness score of 2 if

1. The response is a reason for the verity of the proverb as literally stated or is an elaboration of its meaning and the explanation or elaboration is based on either physical attributes of the symbols or associates to the symbols in the proverb, for example, "Rome is a big city."

2. The response is yes or no or an equivalent response.

3. Both halves of the proverb receive a literalness score of 1 by the criteria listed below.

When the response is scorable, one half is scored 1 for literalness if

1. The response half includes a repetition of a symbol or symbols from the proverb half, for example, "Rome took a long time to complete." *Rome* is a repetition of a symbol. *Took a long time to complete* is an appropriate abstract response for the proverb half. The total literalness score is 1.

2. A synonym for a symbol or a rewording of a symbol from the proverb half is included in the response, for example, "The capital of Italy took a long time." *Capital of Italy* is a synonym for the symbol *Rome*. The total literalness score is 1.

3. The response half includes physical attributes of a symbol from the proverb half, for example, "A big city can't be built in a day."

A *big city* states physical attributes of the symbol *Rome*. *Built in a day* is a repetition of a symbol. Both halves earn a literalness score of 1. The total literalness score for the proverb is 2.

4. The response half is primarily a semantic associate to a symbol from the proverb half, for example, "It took more than one day to build Paris." *Paris* is a semantic associate to *Rome*. *It took more than one day* is a rewording of *not built in a day*. The total literalness score is 2.

A scorable response meeting none of the criteria for literalness receives a literalness score of 0. Responses that are scored 0 include the following: (a) a response that is correct (abstract) according to the Gorham scoring manual; (b) another proverb that has the same meaning as the original proverb; and (c) an attempt at an abstract interpretation of the proverb, even though incorrect; for example, "Big projects require great will power."

The total literalness score for a proverb is the sum of the scores for the two halves. If one half is unscorable, the total score is the score of the scorable half.

The divisions of each proverb into halves and the designated symbols of each proverb are as follows: Form I: 2. Rome/was not built in a day. Symbols: Rome, built, day. 3. When the cat's away/the mice will play. Symbols: cat, mice. 4. Barking dogs/seldom bite. Symbols: Barking, dogs, bite. 5. A stream/cannot rise higher than its source. Symbols: stream, source. 6. Don't swap horses/when crossing a stream. Symbols: horses, stream. 7. The used key/is always bright. Symbols: key, bright. 8. Gold goes in/at any gate except heaven's. Symbols: gold, gate. 9. One swallow/doesn't make a summer. Symbols: swallow, summer. 12. Don't cast pearls/before swine. Symbols: pearls, swine. Form II: 1. He who stumbles twice over one stone/deserves to break his shins. Symbols: stumbles, stone, break, shins. 2. Don't judge a book/by its cover. Symbols: book, cover. 3. The proof of the pudding/is in the eating. Symbols: pudding, eating. 4. One may ride a free horse/to death. Symbols: ride, horse, death. 5. A rolling stone/gathers no moss. Symbols: rolling, stone, moss. 6. Strike/while the iron is hot. Symbols: strike, iron, hot. 7. All is not gold/that glitters. Symbols: gold,

Table 1

Mean Literalness, Concreteness, and Abstraction Scores on 15 Proverbs

Variable	Schizophrenics	Normal subjects
Gorham's concreteness ^a	9.68	2.10
Literalness	7.65	1.34
Gorham's abstraction	6.53	17.51

^a The concreteness values for the (0, 1) Gorham system have been doubled to make them comparable to the (0, 1, 2) values for literalness.

glitters. 10. Let sleeping dogs/lie. Symbols: sleeping, dogs, lie. Form III: 1. The sun/shines upon all alike. Symbols: sun, shines. 2. The grass is always greener/in the other fellow's yard. Symbols: grass, greener, yard. 5. A drowning man/will clutch at a straw. Symbols: drowning, straw. 6. Too many cooks/spoil the broth. Symbols: cooks, broth. 7. The worst spoke/in the wheel breaks first. Symbols: spoke, wheel. 9. It never rains/but it pours. Symbols: rains, pours. 11. There's many a slip twixt the cup/and the lip. Symbols: cup, lip.

Use of the Scoring Scheme with Clinical Groups

Forms II and III of the Gorham Proverbs Test were administered to 77 schizophrenics and 38 firefighters. A brief verbal IQ test consisting of the Comprehension, Vocabulary, and Similarities subtests of the Wechsler Adult Intelligence Scale was also given to these schizophrenics. The firefighters cannot be viewed as control subjects for the schizophrenics because of the lack of full information on their demographic characteristics. The firefighters' data do, however, provide some information on literalness scores of normal subjects.

Mean age of the schizophrenic sample was 37.3 years ($SD = 10.2$), mean years of education was 11.7 ($SD = 3.3$), and mean score on the Hollingshead Index of Social Position was 47.2 ($SD = 15.2$). Mean prorated verbal IQ on the brief intelligence test was 92.5 ($SD = 18.0$). Sixty-two percent of the sample was male, 38% was female. Ninety-five percent was white, 5% was black. Mean score on the Phillips Scale of Premorbid Adjustment was

17.7 ($SD = 4.4$). Mean length of hospitalization was 109.6 months ($SD = 109.9$).

All firefighters were white males. Assuming that the average firefighter receives a high school diploma, their average score on the Hollingshead index would be 51.0. No information on the age, IQ, or marital status of the firefighters was available.

The Proverbs Test was administered using Gorham's instructions, and the responses were scored for literalness by the first author and for concreteness using Gorham's criteria by a graduate student. The scorers were kept blind as to whether a protocol was that of a schizophrenic or a normal subject. To assess interrater reliability, a third scorer rated 40 schizophrenics' protocols according to both systems. To assess the relationship of adequacy of proverb interpretation to both concreteness and literalness, the graduate student also scored all protocols for abstraction using Gorham's manual.

Reliability

The coefficient alpha estimate of reliability for the 115 subjects was .85 for literalness, .84 for Gorham's concreteness, and .92 for Gorham's abstraction. The corresponding values for the 77 schizophrenics were .82 for literalness, .81 for Gorham's concreteness, and .92 for Gorham's abstraction. The correlation between concreteness and literalness was .80 for both groups combined and .74 for the schizophrenics. Interrater reliability for the 40 subjects was .90 for both concreteness and literalness.

Relation to Clinical Status

Table 1 gives the mean scores of both groups according to both scoring systems. As seen in Table 1, both groups received lower scores on literalness than on Gorham's concreteness. Schizophrenics were significantly different from normal subjects on literalness, concreteness, and abstraction ($p < .001$, in each case).

Relation to Intelligence and to Abstraction Score

For the schizophrenics, Verbal IQ correlated $-.52$ ($p < .01$) with concreteness but only

-.15 (ns) with literalness. Thus literalness is less affected than concreteness by sheer intellectual ability. We interpret these values to mean that literalness is less affected than concreteness by generalized deficit. The relation of score on abstraction to literalness and to concreteness lends further support to this interpretation. For the schizophrenics, abstraction correlated $-.64$ with Gorham concreteness and $-.48$ with literalness. Thus abstraction accounted for 41% of the variance of concreteness but only 23% of the variance of literalness. Both correlations were inflated by the fact that concreteness and literalness were obtained from the same responses as abstraction, but this artifact should not affect the correlation for literalness any differently than the correlation for concreteness. The difference between the two correlation coefficients was significant, $t(74) = 2.61, p < .01$, as indicated by a t test for correlations based on dependent data. For the combined group of normal and schizophrenic subjects, abstraction correlated $-.73$ with concreteness and $-.62$ with literalness. Thus abstraction accounted for 53% of the variance of concreteness and 38% of the variance of literalness. The difference between the two correlation coefficients was, again, significant, $t(112) = 2.65, p < .01$.

Conclusion

Literalness of proverb interpretation is less affected by intelligence and by ability to respond abstractly than is Gorham's concreteness. Concreteness depends too much on generalized intellectual deficit to be maximally useful for describing schizophrenic thought disorder. Literalness is an important and more specific kind of error than concreteness and is less affected than concreteness by generalized deficit. The present scoring scheme for literalness should be useful in many situations in which Gorham's concreteness has been used in the past.

Reference Note

1. Friedes, D., Grisell, J. L., Levin, S., Dobie, S., & Cohen, D. B. *Manual for scoring proverb interpretation*. Unpublished manuscript, Lafayette Clinic, 1964. (Available from James Grisell, Lafayette Clinic, 951 East Lafayette, Detroit, Michigan 48207)

References

- Gorham, D. R. *Clinical manual for the Proverbs Test*. Missoula, Mont.: Psychological Test Specialists, 1956. (a)
 Gorham, D. R. A proverbs test for clinical and experimental use. *Psychological Reports*, 1956, 2, 1-12. (b)

Received April 11, 1977 ■

Case Study

Cohesive and Dispersal Behaviors: Two Classes of Concomitant Change in Psychotherapy

Leonard M. Horowitz
Stanford University

Harold Sampson and Ellen Y. Siegelman
Department of Psychiatry
Mt. Zion Hospital and Medical Center
San Francisco

Joseph Weiss
San Francisco Psychoanalytic Institute

Shirley Goodfriend
University of California, Berkeley

This article differentiates between two important classes of behavior that can be identified in any psychotherapy. One class concerns *cohesive* behaviors (Type C), which bring organisms together, and the other concerns *dispersal* behaviors (Type D), which drive organisms apart. This study examined changes in C and D behaviors that occurred during the first 100 hours of the psychoanalytic treatment of Mrs. C, a woman whose presenting complaint was sexual frigidity. The data showed improvement in both types of behaviors. In addition, progress in Type D behavior preceded progress in Type C behavior, a relationship that had been predicted by the case formulation. Then we identified approximately 350 complaints made by the patient during the treatment, complaints of the form "I can't (do something)" and "I have to (do something)." These complaints also declined in frequency during the treatment.

Personality theorists (e.g., Horney, 1945; Murray, 1938) have sometimes classified interpersonal behaviors into three broad categories. This classification is generated by considering, first, whether the subject is (a) avoiding the other person or (b) getting involved with the other person. If the latter, the behavior can be further classified into

(b₁) behaviors expressing a positive involvement and (b₂) behaviors expressing a negative involvement. The resulting three categories can be labeled (a) *avoidance*, (b₁) *positive involvement*, and (b₂) *negative involvement*. Horney (1945) has called these three classes *moving away from*, *moving toward*, and *moving against* the other person. Murray (1938) has written of *abience*, *adience*, and *contrice* with similar meaning.

This study was supported in part by Grant MH 13914 of the National Institute of Mental Health. The authors would like to thank the many trainees and staff members of the Mt. Zion Psychiatric Clinic who generously supplied the clinical judgments and ratings.

Requests for reprints should be sent to Leonard M. Horowitz, Department of Psychology, Stanford University, Stanford, California 94305.

The results reported here are conceptualized in more strictly psychoanalytic terms in Bulletin No. 2 of the Psychotherapy Research Group, Mt. Zion Hospital and Medical Center, by Joseph Weiss and Harold Sampson. A copy of this bulletin can be obtained from Harold Sampson, Department of Psychiatry, Mt. Zion Hospital and Medical Center, P. O. Box 7921, San Francisco, California 94120.

Behaviors involving other people, both positively and negatively, are frequently examined in psychotherapy. Those reflecting positive involvement occur when a person cooperates, collaborates, or concurs with another person, complies and shares thoughts and feelings and is intimate, warm, and loving. Ethologists (Mussen & Rosenzweig, 1973, chap. 28) have called these positive behaviors *cohesive* behaviors, since they bring organisms together. Since many cohesive behaviors begin with the letter *c* (cooperate, concur, comply), we shall call them Type C behaviors.

In contrast, behaviors reflecting negative involvement produce a psychological differentiation from the other person. They occur when a person

defines another person, disagrees with, distrusts, or disapproves of the other person, hates, criticizes or opposes the other person. Ethologists have called these behaviors *dispersal* behaviors, since they (assertively and aggressively) drive organisms apart. Since many dispersal behaviors begin with the letter *d* (defy, disagree, disapprove), we shall call them Type D behaviors. As ethologists have noted, C and D behaviors show a complex interplay throughout the phylogenetic scale, promoting the survival of both the individual and the species.

Frequently during psychotherapy people complain of having poor control over C and D behaviors. Either they are unable to express the behavior comfortably or they are unable to modulate the behavior. In the case described below, for example, the woman sometimes wanted to be affectionate but found herself provoking. At other times she wanted to demur but found herself yielding. Her poor control was accompanied by psychological distress.

Impulse Versus Behavior

To clarify poorly controlled behaviors, let us begin with a basic postulate of psychoanalytic (as well as other) theories, namely that an impulse¹ precedes any nonreflexive behavior. The distinction is analogous to the psycholinguist's distinction between the underlying abstract representation of a thought and the corresponding surface structure of verbal behavior; one, an inferred private event, precedes the other, an observable surface phenomenon. The impulse, an encoded representation, becomes decoded through a grammar that involves optional and obligatory rules and transformations; the defense mechanisms would thus be viewed as a subset of transformations that occur during decoding (cf. Suppes & Warren, 1975).

Just as the correspondence between the deep and the surface structure of language is not necessarily 1:1, the impulse is not necessarily isomorphic to the behavior; different relationships can exist between them. Sometimes an impulse is directly expressed in behavior, at other times the behavior is simply inhibited, and at still other times, the behavior is partly camouflaged by another behavior derived from another impulse. Thus, if a behavior were to exhibit both a C and a D component, we would assume that two different impulses, a C and a D impulse, both existed. An affectionate pinch, according to the postulate, would result from simultaneous impulses to hurt and to be close to the same person.

A psychological "problem" is experienced when

people lack control in translating impulse into behavior. For example, they might intend to express one impulse and yet find themselves expressing another coexisting impulse. That is, on the one hand, they might find themselves unable to express an intended behavior directly and complain, for example, that they cannot cooperate or cannot fight even though they want to. On the other hand, they might find themselves expressing a behavior more intensely or more compulsively than they want to, complaining that they *have* to share intimacies or *have* to defy even though they do not want to; such behaviors would have an obligatory quality.

A successful therapy should help people gain control over each kind of behavior. They should acquire the capacity to experience and express more directly both C and D behaviors. One goal of the following studies is to objectify such improvements and to examine the relationship between them.

Observation 1: Two Concomitant Changes

Method

This set of studies was based on a psychoanalytic case treated by a psychoanalyst who was not familiar with the views expressed here. Every session of the analysis was tape-recorded with the written consent of the patient. The analyst also took process notes during each hour describing the content of the hour. As the patient was talking, the analyst was writing. His notes, however, did not report any commentary or clinical inference; they only summarized the patient's talk and his own interventions.

A group of clinical psychologists and psychoanalysts met weekly to discuss the case. Drawing only on the process notes of the first 10 hours and information of the intake interview, they formulated the case and predicted a sequence of changes. The following case description summarizes the main details of the case and the group's formulation and clinical prediction.

Case description and formulation. The patient, Mrs. C, was a prim, married schoolteacher in her late 20s who came to treatment complaining of sexual frigidity, difficulty experiencing pleasurable feelings, and low self-esteem. Her father was a professional man, and her mother was a housewife. She was the second of four children

¹ The term *impulse* is meant to be neutral theoretically in the way that the term *underlying abstract representation* is neutral. Thus, for example, no energetic connotations are intended.

(an older sister, a younger sister, and a much younger brother). When the treatment began, Mrs. C had been married for less than 2 years. She considered her marriage successful, though she felt that her sexual inadequacy created a major marital problem.

Mrs. C's parents were described as controlled people, undemonstrative of any affection. The mother, who was an organized and efficient woman, ran the house well. She was also very controlling, and Mrs. C felt in danger of being "owned" by her. On the other hand, the mother was not able to defend herself very well. Once, for example, the patient hit the mother in the stomach, and the mother could not defend herself or correct the patient except by retiring to her bedroom in obvious discomfort, leaving the patient to feel guilty, helpless, and frightened. The patient thus came to feel capable of hurting other people and guilty over aggression and assertiveness. Between the ages of 5 and 8, she had recurrent nightmares of something happening to her mother.

The father was also undemonstrative and easily embarrassed by other people's display of affection. Although he was generally controlled, he sometimes lost control of his anger and had temper tantrums that revealed murderous rage; at times Mrs. C felt that he was capable of killing her. The father was also upset by crying women and became angry over masochistic displays from the patient.

In the period before the analysis began, Mrs. C was feeling beleaguered and upset. In situations that called for intimacy, she experienced intense ambivalence, which left her feeling confused and in turmoil. The ambivalence resulted from numerous opposing tendencies: If she had an impulse to be sadistic, for example, she felt potentially guilty. Then, identifying with her mother, she would turn to masochistic feelings (feeling hurt, victimized, neglected, unfavored), which served to camouflage sadistic impulses. These feelings, however, were also unsatisfactory in that she felt that they would upset other people as they had upset her father, who sometimes lost control of his sadism. Thus, she could not express either sadistic or masochistic impulses comfortably and shifted between them, using each tendency to undo the other. The result was turmoil and confusion, and she was sometimes unable to focus her thoughts. She also felt vulnerable to criticism, since she was unable to defend herself against others.

Mrs. C's sexual frigidity may be related to her difficulties with aggression: Since she could not comfortably disengage herself from other

people, sexual intimacy could be a problem in that she did not have the means of ending the closeness when she wanted to. Thus, an impairment in Type D behaviors could produce a corresponding impairment in Type C behaviors.

It was therefore hypothesized that during the treatment, Mrs. C first had to develop a better capacity to defend herself against other people (stubbornly resist other people without feeling guilty, disagree with other people, etc.) to allow herself to get closer to other people. It was hypothesized that as she acquired a better capacity for Type D behavior, she would feel less vulnerable in expressing Type C behavior and would therefore express Type C behaviors more freely.² The following procedure was designed to test these hypotheses.

Procedure. The first step was to examine Mrs. C's ability to express behaviors of Type D comfortably. A prominent subset of Type D behaviors contained instances in which she blamed, criticized, disagreed with, or opposed another person (the therapist or someone else). Three clinicians independently read the process notes of the first 100 hours in unsystematic order to avoid bias, looking for all passages in the notes that described such behaviors (in the present or in the past, toward the therapist or anyone else). Then the three clinicians together reviewed all of the passages that they had identified and retained the ones that they agreed were instances of blaming, criticizing, disagreeing, or opposing. Their resulting set contained 190 passages, involving both direct behaviors (e.g., criticizing the analyst) and self-reports of such behaviors.

Then a 4-point rating scale was developed to assess the directness of the behavior described in each passage. If the blame or criticism was only implied or if it was expressed tentatively with extreme discomfort, the scale value was 1; as the behavior became more explicit and direct, the scale value increased. A rating of 2 meant that a Type D behavior was expressed but immediately undone. Ratings of 3 and 4 meant that a Type D behavior was overtly expressed—3 by telling a third party about it, and 4 by directly

² It might be noted at this point that during the first 100 hours of treatment (covering a period of approximately 6 months), Mrs. C did achieve an increased, though limited, capacity to respond sexually. She also became more able to free associate easily, to reveal symptoms and preoccupations, and to think and work more productively. She became more able to express and tolerate strong feelings and found herself exercising a better-modulated discipline over the students in her class.

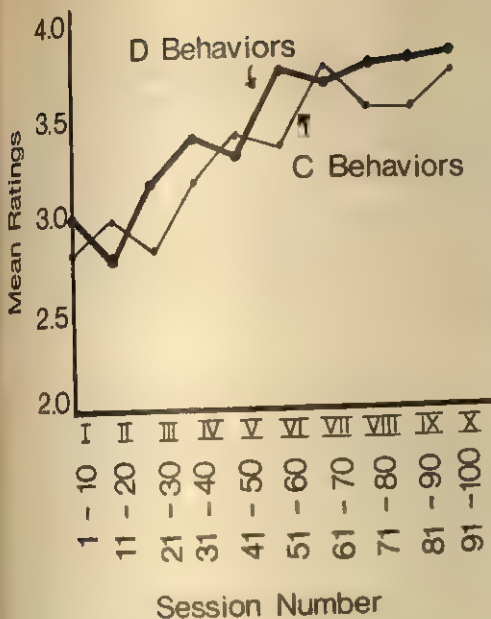


Figure 1. Mean rating of passages in each block of hours.

confronting the offending party. Each rating was also increased by .5 if the event occurred in the present tense (after the treatment began). Thus, the possible ratings were 1.0, 1.5, 2.0, . . . , 4.5.

The passages were divided into two subsets, and the passages of each subset were presented independently in random order to four clinical psychologists who were naïve about the case. Explicit scoring rules were developed for rating the passages, and the judges followed these rules in rating each behavior.

The second major step was to follow a similar procedure for identifying and rating the Type C behaviors. A prominent subset of these included behaviors in which the patient complimented someone, felt affection or compassion for someone, or wanted to be loved by someone. Two clinical psychologists read the process notes of the first 100 hours in unsystematic order looking for all passages in the notes that described such behaviors. As with the Type D passages, the clinicians reviewed the passages that they had identified and retained the ones that they agreed were instances of Type C behaviors. The resulting set contained 106 passages.

A 4-point rating scale was also developed to assess the directness of these behaviors. If the closeness was implied or expressed with extreme uncertainty, lack of clarity, or discomfort, the scale value was 1. A rating of 2 denoted an expression of closeness with immediate undoing.

As the behavior became more explicit and direct, the scale value increased: 3 indicated that the feeling of closeness was expressed to a third party, and 4 indicated that it was expressed directly to the other person. Each rating was also increased by .5 if the event occurred in the present tense (after the treatment began). The possible ratings thus ranged from 1.0 to 4.5.

These passages were also presented to a panel of four clinical psychologists who were naïve about the case. Explicit scoring rules were developed for rating the passages, and the judges followed these rules to rate each behavior.

Results

Type D behaviors. To assess the reliability of the judges' ratings, the four ratings for a given passage were averaged, and the reliability of the four judges' means was computed for each set through an analysis of variance. The reliability was .89 for one set, .90 for the other set.

The 100 sessions were then grouped into 10-session blocks denoted I, II, III, . . . , X. The number of passages within each block were I = 31; II = 13; III = 18; IV = 9; V = 35; VI = 13; VII = 13; VIII = 10; IX = 29; and X = 19. The ratings of passages within each block were averaged, and the means ranged from 2.62 to 3.81. These means are reported in Figure 1, which shows the development of Type D behavior across successive blocks of sessions.

To examine changes in Type D behaviors more closely, all passages rated alike were examined as a group. Because of the small frequencies in some categories, passages rated 2.0 and 2.5 were combined, as were passages rated 4.0 and 4.5. Also, to obtain more stable frequencies, the sessions were grouped into 20-session blocks.

The relative frequency of each rating was computed for each block, and these relative frequencies are shown in Figure 2. The top two graphs show a monotonic decline for passages rated 1.0-2.5. Figure 2 also shows a decline in 3.0s (criticizing someone for a past event) but an increase in 3.5s (criticizing someone for a current event). Direct confrontations (4.0 and 4.5) also became more frequent throughout the treatment. The graphs are largely monotonic and characterize major changes that occurred in the patient's behavior during the treatment.³

³ These graphs, of course, are not independent of one another. The overall improvement in Figure 1 requires that the lower ratings generally decline over the 100 hours while the higher ratings generally increase.

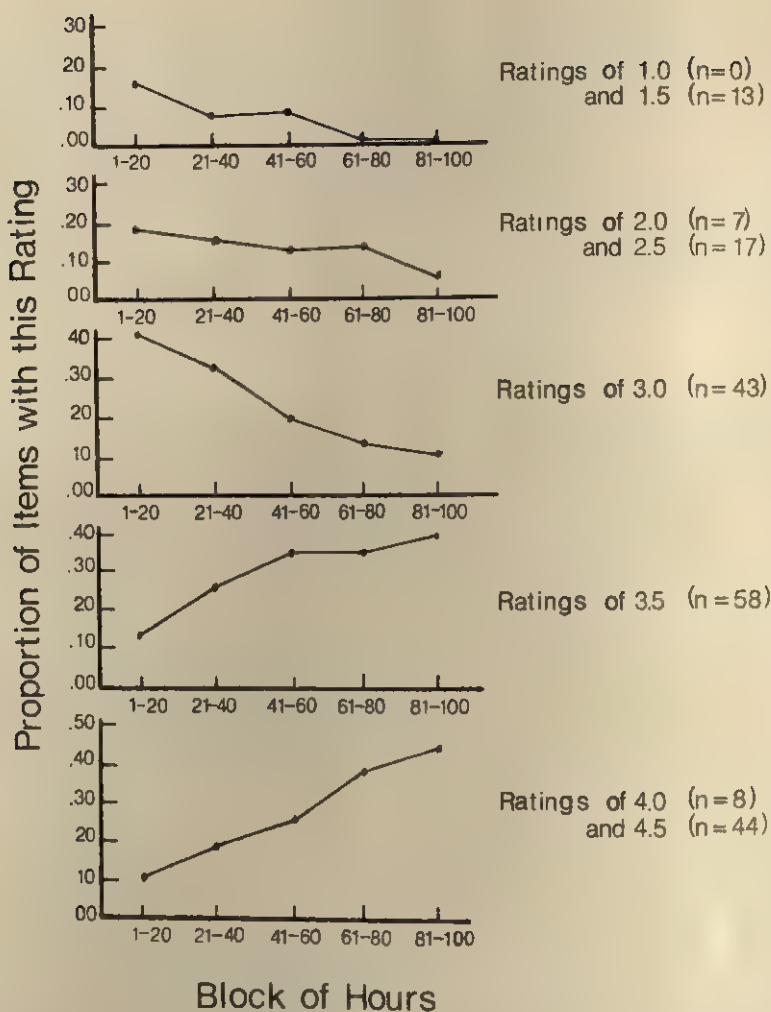


Figure 2. Relative frequency of passages in each rating category for successive blocks of hours (Type D).

To compare the frequencies of past and present events, the relative frequencies of 3.0s and 3.5s were compared. There were 101 passages with these ratings. For each block of 20 sessions, the relative frequency of 3.5s (present tense) was computed. For successive 20-session blocks, the values were 6/24 (i.e., 6 cases were in the present tense, 18 were in the past tense) = .25; 8/16 = .50; 17/26 = .65; 8/11 = .73; and 19/24 = .79. Thus, the patient increasingly came to criticize others for events in her current life. It is assumed that events from the past tense were less threatening for her and provided a convenient starting point for the therapy, but as the sessions progressed, she shifted her focus to her current life. Thus, *part* of the increment in Figure 1 is

due to the patient's shift to present tense events, and *part* is due to the decline in lower ratings (primarily 1.5 and 2.5).

Type C behaviors. The reliability of the 106 Type C ratings was also assessed. The four judges' ratings for a given passage were averaged, and the reliability of the four judges' means was .83. The 100 sessions were grouped into 10-session blocks, with the following frequencies within each block: Block I = 11; II = 9; III = 15; IV = 10; V = 9; VI = 12; VII = 14; VIII = 9; IX = 5; X = 12. The ratings of the passages within each block were then averaged, and the resulting means ranged from 2.67 to 3.75. Figure 1 shows the development of the Type C behaviors across successive blocks of sessions.

To examine the change in Type C behaviors more closely, passages within each rating category were examined separately. Since the frequencies were smaller than those concerning Type D behaviors, all of the ratings from 1.0 to 3.0 were pooled (These were the categories that had shown declining relative frequencies in Type D behaviors.); likewise, all of the ratings from 3.5 to 4.5 were pooled. (These were the categories that had shown increasing relative frequencies in Type D behaviors.) Relative frequencies of occurrence were computed for each block of 20 sessions, as shown in Figure 3. One graph shows a progressive decline in the relative frequencies of the lower ratings, and the other graph shows a progressive increase in the relative frequencies of the higher ratings. The two graphs thus resemble those obtained for the Type D behaviors. Events in the past tense were also examined, but their frequencies were too small (only 23 cases) to permit any inference.

Thus, it is clear that two types of changes occurred, but it still needed to be demonstrated that a change occurred in the patient's presenting complaint, sexual frigidity. Therefore, every reference to the patient's sexual behavior was noted throughout the 100 hours. There were 18 such references (comprising a subset of the 106 Type C passages), all occurring between Hours 28 and 100. Each passage contained the word *intercourse* except one, which contained the phrase *sexual interest*. Here are some examples: From Hour 33 (rated 2.5): "Sometimes when she is trying to make herself have intercourse with Bill, she feels as though she wants to hurt him. She just doesn't understand it. She'll go from feeling very warm to feeling nothing toward him suddenly." From Hour 67 (rated 4.5): "This weekend she and Bill had intercourse, and she was thinking how different it can be when she's thinking about him and feeling close to him and not all wrapped up in herself."

Seven passages occurred in the first 50 sessions, and 11 occurred in the last 50 sessions. The C rating assigned to each passage was noted. For those in the early block, 6 had ratings of 1.5 to 2.5, and 1 (in Hour 43) had a rating of 3.5 to 4.5. Of the 11 passages in the later block, 4 had ratings of 1.5 to 2.5, and 7 had ratings of 3.5 to 4.5. The 7 passages with high ratings were mainly simple, direct statements that the patient had had sexual intercourse. A Fisher exact test was performed to test the significance of this difference; the chance probability of the observed pattern or a more extreme one is .022.

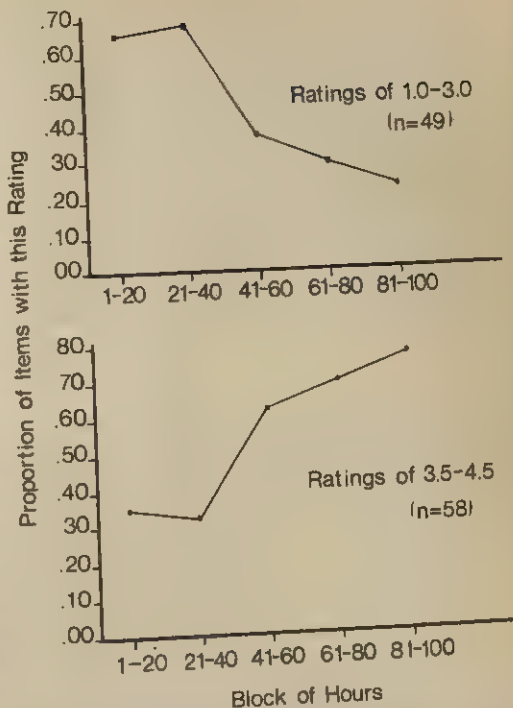


Figure 3. Relative frequency of passages in different rating categories for successive blocks of hours (Type C).

Observation 2: Relationship Between C and D Behaviors

The capacities to express C and D behaviors comfortably seem to be related, since a defect in D can produce a corresponding defect in C. That is, if a person did not have the capacity to disengage from the other person, intimacy would be unsafe, since the person would not be able to end the closeness and would run the danger of feeling oppressed or entrapped. On the other hand, once the person gained the capacity to express D behaviors comfortably, closeness would not be as threatening.

Thus, as Mrs. C gained the capacity to express Type D behaviors comfortably, it should become easier for her to express Type C behaviors. In any block of therapy sessions in which significant gains are observed in Type D behaviors, improvement should subsequently be observed in Type C behaviors. This hypothesis is examined below.

Method, Results, and Discussion

In Figure 1 the C graph resembles the general form of the D graph. To examine the relationship between the graphs more closely, the posi-

tions of greatest increase along each graph were noted. A "significant improvement" in either function is defined as an increment from Block i to Block $i + 1$ that exceeded .25. Significant improvements in Type D behavior occurred three times—from Block II to Block III, from Block III to Block IV, and from Block V to Block VI. Furthermore, significant improvements in Type C behavior also occurred three times—from Block III to Block IV, from Block IV to Block V, and from Block VI to Block VII. In each case, a significant improvement in Type C behavior followed a significant improvement in Type D behavior: An improvement in D occurred from II to III, an improvement in C occurred from III to IV. The chance probability that the three Type C improvements would occur in these particular three positions is .012.

In addition, a "setback" in either function is defined as a decrement from Block i to Block $i + 1$. A setback in the Type D behavior occurred three times—from Block I to Block II, from Block IV to Block V, and from Block VI to Block VII. A setback also occurred three times in the Type C behavior—from Block II to Block III, from Block V to Block VI, and from Block VII to Block VIII. Thus, a setback in Type C behavior always followed a setback in Type D behavior. In other words, the two graphs took very similar courses, with one displaced from the other by one block of sessions.

The data therefore suggest that the patient's progress in expressing Type C behaviors followed her progress in expressing Type D behaviors. As she became progressively able to criticize, oppose, and disagree with other people, she felt progressively less vulnerable; then, feeling less vulnerable, she could relax her defenses and permit herself to feel close, affectionate, and compassionate toward other people. If the two graphs had simply exhibited a correlation, other factors could account for their concomitant rise and fall. But their displacement in time suggests that an advance in one type of behavior may have facilitated an advance in the other.

This inference must be made with reservations for three reasons. First, the relationship may only describe an idiosyncrasy of one patient's progress and needs to be replicated with other cases. For example, Mrs. C's progress in part reflected a shift from past to present tense, and her proportions of present tense passages throughout the treatment, while similar for C and D behaviors, were not identical. It is possible that combining nonuniformities of this type could produce a lag between two graphs. Such issues would be best resolved by replicating the findings on another case.

Second, changes in Type C and Type D behaviors, as operationalized here, may be trivial. That is, they may reflect changes that occur in any developing human relationship in the way that the partners relate to each other (talking more directly, less cautiously, less formally) and are thus not necessarily to be traced to the therapy itself. It is possible that whenever Mrs. C entered a new relationship with someone, she would initially qualify with great caution any statements that she made so as to present a balanced view on any subject; such a tendency would involve statements that would get lower ratings. Then, as she came to know the other person better, she might drop this tendency and become more direct. If this interpretation were correct, though, C and D changes should occur simultaneously, rather than one consistently lagging behind the other.

Finally another kind of explanation might account for the observations in Figure 1. Suppose the direct expression of aggression is in some sense incompatible with the direct expression of intimacy, so that the relative prominence of one would imply a relative decline in the other. Then, as one graph rose from Block i to Block $i + 1$, the other graph would fall. For example, in Figure 1, from Block I to Block II, the C graph rises while the D graph falls, causing the graphs to cross. Then, proceeding to Block III, the C graph falls while the D graph rises, producing another crossing. Additional crossings occur as the graphs proceed to Blocks V, VI, VII, and VIII. This characterization of the data has the virtue of parsimony, but it does not explain why both graphs would show concomitant overall improvement. It also suggests that the frequency of Type D behaviors should be strongly and negatively related to the frequency of Type C behaviors. The correlation was negative, but it was not significant ($r = -.33, p > .20$).

Thus, alternative hypotheses may account for some aspects of the data, and perhaps may even accurately account for aspects of the therapeutic process. However, they do not adequately explain the lag between graphs or the overall improvement in each type of behavior. For this reason, it is tentatively concluded that improvement in Type D behavior, at least in this patient, permitted subsequent improvement in Type C behavior.

Observation 3: The Nature of Mrs. C's Complaints

In the courses of 100 hours of treatment, Mrs. C mentioned a large number of other problems

that were not directly related to sexual frigidity but that clarify the nature of her distress. Many of these complaints were expressed in the form "I can't (do something)" or "I have to (do something)," revealing inhibitions and compulsions. A large subset of these complaints could be classified according to the C and D categories, and it was hypothesized that many complaints would reflect general problems over C and D behaviors.

Two people reading the process notes independently identified 248 complaints involving "can't" (e.g., She can't praise her assistant) and 103 complaints involving "has to" (e.g., She has to fight against her husband), making a total of 351 complaints.⁴ Near symptoms of *can't* and *has to* were also accepted.

The statements were presented to a group of 20 judges (10 graduate students and 10 clinicians). Each judge was asked to classify each problem behavior as to Type C, Type D, or neither. A statement was considered a Type C or Type D complaint if 14 or more judges so classified it. Using this 14-or-more criterion, 60 complaints were of Type C and 56 were of Type D.

The complaints of each type were then examined further to determine how many were of the *can't* form and how many were of the *has to* form. Of the 60 Type C complaints, 51 were of the *can't* form and 9 of the *has to* form. Of the 56 Type D complaints, the corresponding frequencies were 24 and 32. The chi-square computed for this 2×2 matrix was 20.7 ($p < .001$). Whereas Type C complaints were typically of the *can't* form, Type D complaints were more evenly divided between the two. The single highest frequency was for complaints of the form "*can't C*," a form that corresponds to the presenting complaint, sexual frigidity.⁵ The other complaints, involving aggression and assertiveness, reflected poor control both ways: Sometimes the patient could not express behaviors that she wanted to express, but at other times she could not restrain herself.

General Discussion

Sophisticated studies of psychotherapy outcome have been undertaken in recent years, as summarized in the recent review of Bergin and Suinn (1975). Most of these studies (e.g., Berzins, Bednar, & Severy, 1975; Sloane et al., 1975) have reported data about treatment outcome, though details of the therapeutic process remain generally unclear. The present set of studies, in contrast, focuses on the treatment process per se and assumes that therapeutic out-

come is best evaluated in the light of one patient's needs and goals.

The present article has examined several explicit propositions about the nature of Mrs. C's psychopathology and therapeutic progress. One major result showed that Mrs. C's difficulty in expressing Type C behavior was related to her difficulty in expressing Type D behavior; thus, the way to solve one specifiable set of problems involved the simultaneous treatment of another set. Throughout the treatment, progress on one set was a prerequisite for concomitant progress on the other.

These results thus point out one feature of a therapeutic process that is often overlooked in treatments that set more specific behavioral goals, namely, that an advance in one behavioral domain may be a prerequisite for an advance in another, quite different, domain. For a patient like Mrs. C, a gain in assertiveness may be necessary for a gain in intimacy. Occasional writers have implied such a relationship (e.g., Smith, 1975), but no systematic documentation or explanation of the relationship has previously been offered.

Furthermore, in Mrs. C's treatment, there were really two major therapeutic goals, but only one corresponded to her presenting complaint (sexual frigidity). It is possible, of course, that Mrs. C would have been helped more efficiently by a combination of assertiveness training and sexual therapy, but it is not necessarily the case that she perceived herself as needing to become more assertive. Indeed, a tabulation of her complaints throughout the first 100 hours showed that she often found herself too aggressive and oppositional, more than she wanted to be. Nonetheless, in principle, one could imagine a research design with patients like Mrs. C, comparing each kind

⁴ These various complaints declined in frequency over the 100 hours. The relative frequencies occurring in successive 20-session blocks were .26, .24, .17, .19, .15; $\chi^2(4) = 15.48$, $p < .01$.

⁵ Very few of these *can't C* complaints were specifically sexual in content, however. They concerned various people—the patient's husband, therapist, pupils, assistants—and they involved various forms of closeness—giving unilaterally to other people (praising, helping, reassuring, comforting, disclosing personal information), as well as exchanging (relating to other people, trusting, believing, returning love, feeling close to or relaxed with). The Type D problem behaviors were also heterogeneous, involving assertiveness (getting her own way, sticking to her views, developing her own teaching method, making demands on other people, disagreeing with other people) and aggression (expressing anger, being nasty to people, criticizing other people, opposing other people).

of treatment singly and in combination with the other kind of treatment.

One early theme in Mrs. C's therapy consisted of her criticizing people (e.g., her parents) for events of the past. This kind of theme often occurs early in a treatment as the patient spontaneously produces data from the past. It is possible that Mrs. C saw as one demand characteristic of therapy that she criticize her parents for events of the past. In our view, however, she was not only producing personal data but was also serving very specific therapeutic ends by beginning the treatment in this way. Her criticisms allowed her to observe the therapist's reaction to one very mild form of aggression and assure herself of the safety of similar undertakings in the future. This low-level criticizing can be viewed as an early test of a therapist. Other evidence of such tests has been presented by Horowitz, Sampson, Siegelman, Wolfson, and Weiss (1975).

Finally, the distinction between C and D behaviors emphasizes the *meaning* of the behavior in addition to its observable form. A given behavior may have multiple meanings for a particular person. For example, the very same behavior might be of Type D with respect to one person and of Type C with respect to another person. A criticism rated 3 in this study would be such a case; it was of Type D with respect to the criticized person and of Type C with respect to the therapist (since the patient is confiding in or confessing to the therapist). This form of closeness to the therapist was never tabulated among the Type C behaviors of this study, but it may comprise a significant aspect of the

therapeutic process: The patient criticizes a third person, tentatively viewing the therapist as an ally; then, when the therapist permits the alliance, a closeness is established between them that neither party has directly solicited. Such aspects of the therapeutic process need to be examined further.

References

- Bergin, A. E., & Suinn, R. M. Individual psychotherapy and behavior therapy. *Annual Review of Psychology*, 1975, 26, 509-556
- Berzins, J. L., Bednar, R. L., & Severy, L. J. The problems of intersource consensus in measuring therapeutic outcomes: New data and multivariate perspectives. *Journal of Abnormal Psychology*, 1975, 84, 10-19.
- Horney, K. *Our inner conflicts*. New York: Norton, 1945.
- Horowitz, L., Sampson, H., Siegelman, A., Wolfson, A., & Weiss, J. On the identification of warded off mental contents. *Journal of Abnormal Psychology*, 1975, 84, 545-558.
- Murray, H. A. *Explorations in personality*. New York: Oxford University Press, 1938.
- Mussen, P., & Rosenzweig, M. R. *Psychology, an introduction*. Lexington, Mass.: Heath, 1973.
- Sloane, R. B., et al. *Short-term analytically oriented psychotherapy vs. behavior therapy*. Cambridge, Mass.: Harvard University Press, 1975.
- Smith, M. J. *When I say no, I feel guilty*. New York: Bantam Books, 1975.
- Suppes, P., & Warren, H. On the generation and classification of defense mechanisms. *International Journal of Psychoanalysis*, 1975, 56, 405-414.

Received July 22, 1977 ■

Brief Reports

Locus of Control, Prediction, and Performance on University Examinations

Timothy M. Gilmor and David W. Reid
York University, Downsview, Canada

Fifty-two internal and external locus of control subjects estimated what their outcomes would be for two term exams and their final grade. Internals' estimates and actual outcomes tended to be higher than those of externals. Accuracy of predictions as assessed by difference scores did not differentiate the two groups. However, internals' estimates from the first to second exam were characterized by more typical expectancy shifts demonstrating a greater responsiveness to their initial performance feedback.

Research on the locus of control construct (Rotter, 1966) has shown that internals compared to externals obtain higher grades (Brown & Strickland, 1972) and make more accurate predictions for self-relevant achievement outcomes (Steger, Simmons, & Lavelle, 1973; Wolfe, 1972). However, the former relationship has been found more consistently for younger samples than for older samples (Gilmor, 1978; Phares, 1973), and some research demonstrating the latter relationship has been criticized on methodological grounds (Wolfe & Egelston, 1973).

The present study extends prior work by examining both the actual performance of internal and external university students and the accuracy of their predictions over several term exams as well as the final grade. A more thorough examination of the relationship was possible by broadening the sample of behavior (three outcome events rather than one) and by incorporating additional dependent measures assessing the nature of internals' and externals' predictions. In addition to assessing the accuracy of predictions by means of difference scores, the percentage of overestimates versus underestimates and typical versus atypical expectancy (estimate) shifts from the first to second exam were calculated.

Prior research (cf. Strickland, 1977) has found

that internals exhibit more typical expectancy shifts (increase in expectancy after success, decrease in expectancy after failure), whereas externals exhibit more atypical expectancy shifts (increase in expectancy after failure, decrease in expectancy after success). This pattern of expectancy shifts was expected to occur in the present context and to provide support for the contention that internals are more responsive to the feedback that they receive in making predictions for future performance. No specific prediction was made concerning the proportion of overestimates versus underestimates for each group.

Subjects were 20 male and 32 female undergraduates in a third-year psychology course. Several weeks after the class had begun, the first of several questionnaires was administered. Information sought included: "How many university exams or term tests have you written?" "What is the final grade you expect to get in this course?" and, "It is early to predict, but how well do you think you will do on the first exam in this course? ____/100."

One week later the class completed an expanded version of the Reid and Ware (1974) multidimensional Locus of Control scale. A median split of scores on the Fatalism subscale determined the internal (0-6) and external (8-18) groups. Subjects with scores at the median were eliminated.

After the first exam results were made known to the class, a "follow-up" questionnaire was administered asking subjects to report their exam result and to predict what their second exam result would be (scheduled for 4 weeks later). At the end of the term, their second exam results and

Completion of this study was facilitated by a Canada Council doctoral fellowship granted to the first author.

Requests for reprints and for an extended report of this study should be sent to Timothy M. Gilmor, Department of Psychology, York University, 4700 Keele Street, Downsview, Ontario, Canada M3J 1P3.

final grade were obtained for comparison to their predictions.

An initial analysis revealed no main or interaction effects due to high versus low exam experience (cf. Wolfe, 1972). Overall, the group had considerable previous exam writing experience ($M = 14.46$). Consequently, 2×2 (Sex \times Fatalism) analyses of variance were conducted on the (a) performance estimates, (b) actual results, and (c) difference scores for each exam and the final grade. Chi-square analyses were conducted on the frequency of overestimates versus underestimates and typical versus atypical expectancy shifts.

For the first exam, the estimates of internals ($M = 75.9$) were marginally higher than the estimates of externals ($M = 72.6$), $F(1, 48) = 2.56$, $p < .15$. The actual first-exam performance of internals ($M = 76.9$) was significantly higher than that of externals ($M = 64.0$), $F(1, 48) = 11.02$, $p < .01$. For the second exam, a significant difference was found between internals' ($M = 77.1$) and externals' ($M = 72.0$) estimates, $F(1, 48) = 6.54$, $p < .05$. However, only a marginal difference, $F(1, 48) = 2.40$, $p < .15$, was obtained for internals' ($M = 66.1$) versus externals' ($M = 60.6$) actual performance.

With respect to the final grade, no differences were found between the different groups' estimates, $F(1, 48) < 1.00$, $p > .25$, although internals ($M = 4.0$) compared to externals ($M = 5.1$) obtained higher final grades, $F(1, 48) = 5.00$, $p < .05$. The numerical means approximate a grade letter of B for internals and C for externals. Both groups had initially estimated that they would achieve a grade of B+.

Analyses of the difference scores for each exam and the final grade revealed no sex or locus of control differences.

Analysis of the percentage of overestimates versus underestimates for both exams combined revealed that externals (78.6%) more frequently overestimated their upcoming exam result than did internals (60.4%), $\chi^2(45) = 4.92$, $p < .05$.

Consistent with the above results and with prior research was the finding that the expectancy shifts of internals (62.5%) were more frequently typical (vs. atypical) than those of externals (42.8%), $\chi^2(42) = 3.98$, $p < .05$. That is, internals more frequently raised their estimates for the second exam if expectancies for the first exam had been surpassed, and they lowered their estimates for the second exam if expectancies

for the first exam had not been met. Externals did the opposite; they more frequently raised their expectancies for the second exam despite poorer-than-expected performance on the first exam and lowered their expectancies for the second exam despite better-than-expected performance on the first exam. These results support the theoretical contention that externals are not as responsive as internals to initial feedback in making estimates for future performance.

Although internals and externals in this sample were not differentiated in terms of the accuracy of their predictions, their predictions did differ qualitatively in terms of the frequency of overestimations and typical expectancy shifts. Also, the finding that internals attained higher performance outcomes suggests that the locus of control/academic achievement relationship so often found in studies with children also holds for adult samples.

References

- Brown, J. C., & Strickland, B. R. Belief in internal-external control of reinforcement and participation in college activities. *Journal of Consulting and Clinical Psychology*, 1972, 38, 148.
- Gilmor, T. M. Locus of control and adaptive behaviour in children and adolescents. *Canadian Psychological Review*, 1978, 19, 1-26.
- Phares, E. J. *Locus of control: A personality determinant of behavior*. Morristown, N.J.: General Learning Press, 1973.
- Reid, D. W., & Ware, E. E. Multidimensionality of internal versus external control: Addition of a third dimension and nondistinction of self versus others. *Canadian Journal of Behavioural Science*, 1974, 6, 131-142.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Steger, J. A., Simmons, W. L., & Lavelle, S. Accuracy of prediction of own performance as a function of locus of control. *Psychological Reports*, 1973, 33, 59-62.
- Strickland, B. R. Internal-external control of reinforcement. In T. Blass (Ed.), *Personality variables and social behavior*. Hillsdale, N.J.: Erlbaum, 1977.
- Wolfe, R. N. Perceived locus of control and prediction of own academic performance. *Journal of Consulting and Clinical Psychology*, 1972, 38, 80-83.
- Wolfe, R. N., & Egelston, R. Critique of "Accuracy of prediction of own performance as a function of locus of control." *Psychological Reports*, 1973, 33, 305-306.

Received March 28, 1977 ■

Person \times Situation Interaction in Personality Prediction: Some Specifics of the Person Factor

Lawrence J. Breen
University of Manitoba
Winnipeg, Canada

Terry J. Prociuk
University of Manitoba
Winnipeg, Canada

Norman S. Endler
York University
Toronto, Canada

Marilyn Okada
York University
Toronto, Canada

This study attempted to determine if the multidimensionality of trait anxiety (A-Trait), as measured by the S-R Inventory of General Trait Anxiousness, was related to differential personality profiles. Results of stepwise multiple linear regression analyses, using the Internal-External Locus of Control Scale, the California Psychological Inventory, the Fear of Negative Evaluation Scale, and the Interpersonal Trust Scale as predictor measures, confirmed the existence of distinct personality profiles for each of the four facets of A-Trait. These findings attest to the multidimensionality of the S-R Inventory of General Trait Anxiousness and to the importance of specifying the situation in the measurement of anxiety.

Recently, Endler and Okada (1975) discussed the limitations of many existing omnibus measures of anxiety proneness (A-Trait). At the same time these researchers provided data supporting their claim that the S-R Inventory of General Trait Anxiousness offers a more appropriate assessment of A-Trait. This inventory is a multidimensional measure of anxiety (e.g., Endler & Magnusson, 1976; Endler, Magnusson, Ekhammar, & Okada, 1976) that provides indices of anxiety responses to situations involving interpersonal relations, physical danger, ambiguity, and generally innocuous conditions.

The purpose of this study was to determine if the multidimensionality of the Endler-Okada inventory would be reflected in differential personality profiles. These profiles were to be derived from a battery of personality tests administered with this inventory and were expected to be logically and theoretically consistent with the four situational dimensions of anxiety ostensibly assessed by the S-R Inventory of General Trait Anxiousness.

Subjects were 278 senior-year, female nursing students in training at 13 institutions in the province of Manitoba, Canada. The subjects were

administered a personality battery that consisted of the 18 scales of the California Psychological Inventory (CPI; Gough, 1957). The CPI seeks to measure interpersonal facets of personality such as sociability and self-control. They were also given the Internal-External Locus of Control Scale (Rotter, 1966); the Interpersonal Trust Scale (ITS; Rotter, 1967); the Fear of Negative Evaluation scale (FNE), developed by Watson and Friend (1969); and the S-R Inventory of General Trait Anxiousness (Endler & Okada, 1975). Obtained data were subjected to separate stepwise multiple linear regression analyses with the four S-R situation scores used as the criterion variables.

For parsimony of results only those variables that related significantly ($p < .05$) to the S-R situation scores and yielded a multiple correlation change equal to or greater than .01 are discussed. There were additional variables whose statistical contributions were significant but whose practical contribution was negligible (i.e., multiple correlation change $< .01$).¹

With these considerations in mind, it was found that scores on the interpersonal facet of A-Trait (i.e., you are in situations involving interactions with other people) were negatively related to scores on the CPI measures of intellectual efficiency, sociability, self-acceptance, and

The authors would like to thank Helen P. Glass, Director of School of Nursing, University of Manitoba, for making the data collection possible.

Requests for reprints should be sent to Lawrence Breen, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada R3T, 2N2.

¹ Complete data summary tables are available from the first author on request.

good impression ($R = .55$, $p < .05$). High scores on the physical danger facet of A-Trait (i.e., you are in situations where you are about to or may encounter physical danger) were associated with low scores on two measures (CPI scale of good impression and ITS and with high scores on the CPI Sociability scale ($R = .34$, $p < .05$). For subjects expressing anxiety on the ambiguous facet of A-Trait (i.e., you are in a new or strange situation), there was a negative association with their scores on the CPI measures of sociability, capacity for status, and good impression. These same individuals obtained high scores on the FNE scale and low scores on the CPI measure of flexibility ($R = .54$, $p < .05$). For the innocuous facet of A-Trait (i.e., you are involved in your daily routines), the scores were negatively related to scores on the CPI measures of intellectual efficiency and achievement via conformance ($R = .34$, $p < .05$).

Independence among the four scales of the S-R inventory can be inferred from the distinctions among the predictive regression equations and also (of course) from the correlations among the parts. With respect to the latter, the correlation coefficients were low and ranged from $-.04$ to $.21$ for physical danger versus innocuous and ambiguous versus innocuous, respectively.

In summary, the obtained results strongly suggest the existence of separate personality profiles for the four facets of A-Trait. Furthermore, it is evident that these profiles are logically congruent with the type of situation described. For example, it is reasonable to expect that individuals who express anxiety in an interpersonal situa-

tion should score low on measures of sociability, self-acceptance, good-impression, and so on. Similarly, persons who feel anxious in unfamiliar, ambiguous situations might well indicate a lack of flexibility and a concern over social evaluation. In sum, the data tend to form patterns that are consistent with the claim for the multidimensionality of the S-R inventory.

References

- Endler, N. S., & Magnusson, D. Multidimensional aspects of state and trait anxiety: A cross-cultural study of Canadian and Swedish students. In C. D. Spielberger & R. Diaz-Guerrero (Eds.), *Cross-cultural research on anxiety*. Washington, D.C.: Hemisphere, 1976.
- Endler, N. S., Magnusson, D., Ekehammar, B., & Okada, M. The multidimensionality of state and trait anxiety. *Scandinavian Journal of Psychology*, 1976, 17, 71-83.
- Endler, N. S., & Okada, M. A multidimensional measure of trait anxiety: The S-R Inventory of General Trait Anxiousness. *Journal of Consulting and Clinical Psychology*, 1975, 43, 319-329.
- Gough, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Rotter, J. B. A new scale for the measurement of interpersonal trust. *Journal of Personality*, 1967, 35, 661-665.
- Watson, D., & Friend, R. Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, 1969, 33, 448-457.

Received April 4, 1977 ■

Abnormality of Subtest Score Differences on the WISC-R

Richard J. Piotrowski
Allegheny Intermediate Unit #3
Pittsburgh, Pennsylvania

The abnormality of the difference as a method for evaluating the magnitude of differences between pairs of Wechsler Intelligence Scale for Children-Revised subtests is discussed. Generally, abnormal differences at the .05 level of significance range from 6 to 7 scaled score points and 8 to 10 scaled score points at the .01 level. Abnormal Verbal-Performance scale IQ differences are also considered. Such differences averaged 18 IQ points at the .05 level of significance and 24 IQ points at the .01 level. The diagnostic implications of the use of the abnormality of the difference for evaluating subtest score differences are discussed.

The evaluation of differences between subtest scores is a common problem for the users of tests such as the Wechsler Intelligence Scale for Children-Revised (WISC-R). Unfortunately, the magnitude of subtest score differences has too often been evaluated on the basis of clinical intuition or on the recommendations of test publishers without an adequate statistical basis for such recommendations.

When a statistical answer to this problem has been given, it has generally involved computing the standard error of measurement of the difference between subtest pairs. Using this approach, the WISC-R manual (Wechsler, 1974) provides a table of significant subtest score differences based on the average values for all 11 age groups included in the standardization sample. Using the rather weak .15 level of significance, differences of only 2.35-3.45 scaled score points between subtests are considered significant. Using the more conventional .05 and .01 levels of significance and considering age levels separately, it was found that 3-5 points at the .05 level and 4-6 points at the .01 level were generally required for significance (Piotrowski & Grubb, 1976; Sattler, 1974).

Another way of statistically evaluating subtest score differences, the abnormality of the difference, was suggested by Silverstein (1973). While the standard error of measurement of the difference indicates how large a difference must be so that it cannot be attributed to measurement

error, this other method reveals how large of a difference must exist for there to be little chance of it occurring in a normal population (Silverstein, 1973).

The abnormality of the difference (A_d) can be computed using the following formula:

$$A_d = z\sqrt{\sigma_1^2 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2}$$

where σ_1 and σ_2 are the standard deviations of the subtests being compared, r_{12} is the intercorrelation between these two subtests, and z is the probability level that was used.

The present investigation involved computing the abnormality of the difference between all possible pairs of the 12 WISC-R subtests at each of the 11 age levels included in the standardization sample. Differences between Verbal and Performance scale IQs were also examined. The necessary standard deviations and intercorrelations were obtained from the WISC-R manual (Wechsler, 1974).

For a difference between any two WISC-R subtests to be considered "abnormal," a 6- or 7-point difference is generally required at the .05 level of significance, whereas an 8- to 10-point difference is typically required at the .01 level. Abnormal Verbal-Performance scale IQ differences averaged 18 points at the .05 level of significance and 24 points at the .01 level.

When compared to the size of differences generated by the standard error of measurement of the difference, the abnormality of the difference resulted in subtest scaled score differences for similar comparisons that were generally 2 or 3 points higher at the .05 level and 3 or 4 points higher at the .01 level. Abnormal Verbal-Performance scale IQ differences were 5-8 points higher at the .05 level and 8-10 points higher at

Requests for reprints and for an extended report of this study should be sent to Richard J. Piotrowski, Allegheny Intermediate Unit #3, Suite 1300, 2 Allegheny Center, Pittsburgh, Pennsylvania 15212.

the .01 level than when the weaker standard error of measurement of the difference was used.

In light of the rather large discrepancy in the size of "differences" generated by these two indices, the implications of differences of varying magnitude should be considered. Test score differences falling short of the standard error of measurement of the difference may, of course, be attributed to measurement error or chance. A difference between two scores, which is significant in light of the standard error of measurement of the difference, suggests that there is a difference that can be reliably measured in the abilities or skills tapped by the tests being compared. Such differences may well have educational or program planning significance, since they point to "real" differences between the abilities measured by the subtests being compared. However, the fact that an individual's abilities are not uniformly developed should not be considered highly unusual. Only when the difference between scores reaches the magnitude of the abnormality of the difference might the scatter between an individual's abilities be considered unusual when compared to the amount of scatter generally found within the abilities of *other* individuals.

Two recent studies are relevant in regard to the size of differences obtained when using the WISC-R. Using data computed from the standardization sample of the WISC-R, Kaufman (1976a) found that rather large differences between subtest scores are typical. When all 12 WISC-R subtests were given, less than 30% of the normative population showed a difference of 6 or less points between their highest and lowest subtest scaled scores. At the same time, approximately 20% had differences of 9 points or more, 5% had differences of 11 points or more, and 2% had differences of 13 points or more.

A second study by Kaufman (1976b) revealed that rather large Verbal-Performance IQ scale differences were also typical of the standardization sample. Verbal-Performance IQ differences of 9 points or greater were found in 48% of subjects, 34% had differences of 12 points or more, and 25% had differences of 15 points or more.

At the .05 level, abnormal Verbal-Performance differences averaging 18 IQ points were found in the present study. Differences of this size or greater were found by Kaufman (1976b) in 12% of the standardization samples. Similarly, at the .01 level, abnormal Verbal-Performance differences averaged 24 points. Only 4% of the original sample had differences this large or greater.

In light of Kaufman's findings, it appears that the abnormality of the difference provides a statistic that is realistically related to the actual performance of children on the WISC-R. This close degree of correspondence between the abnormality of the difference and Kaufman's findings has important diagnostic implications. In regard to the identification of children who are expected to show a wide variability between their "peaks" and "valleys" (e.g., learning disabled), differences more in line with the magnitude of the abnormality of the difference or greater might be expected. Similarly, the tolerance for the spread of abilities in children seen as potential candidates for programs characterized by a "flat" ability profile (e.g., educable mentally handicapped, slow learners) might be extended to just short of the point generated by the abnormality of the difference.

References

- Kaufman, A. S. A new approach to the interpretation of test scatter on the WISC-R. *Journal of Learning Disabilities*, 1976, 9, 160-167. (a)
- Kaufman, A. S. Verbal-performance IQ discrepancies on the WISC-R. *Journal of Consulting and Clinical Psychology*, 1976, 44, 739-744. (b)
- Piotrowski, R. J., & Grubb, R. D. Significant subtest score differences on the WISC-R. *Journal of School Psychology*, 1976, 14, 202-206.
- Sattler, J. M. *Assessment of children's intelligence* (Rev. ed.). Philadelphia, Pa.: Saunders, 1974.
- Silverstein, A. B. Reliability and abnormality of differences between DTVP subtest scores. *Psychology in the Schools*, 1973, 10, 204-206.
- Wechsler, D. *Wechsler Intelligence Scale for Children-Revised manual*. New York: Psychological Corporation, 1974.

Received April 6, 1977 ■

Correlation of WAIS IQ in 10 Pairs of Brothers

Joseph D. Matarazzo and Arthur N. Wiens
University of Oregon Health Sciences Center

Allen E. Shealy
University of Alabama in Birmingham

Ten pairs of brothers with a mean age of 24 years and a mean of 13 years of education were individually examined with the Wechsler Adult Intelligence Scale some 10 months apart by a highly experienced clinical psychologist who was unaware of the consanguineous relationship. The obtained correlation of .42 for Full Scale IQ is consistent with the median correlation of .49 reported by Erlenmeyer-Kimling and Jarvik in their 1963 review of the world's literature.

The relationship between consanguinity and measured intelligence was the focus of active investigative interest between 1910 and 1950, with the results reported in the world's literature succinctly summarized in a review by Erlenmeyer-Kimling and Jarvik (1963). Inasmuch as the Wechsler scales had not yet been developed during the period when most of these studies were being conducted, we were curious as to whether the post-1940 studies on the relationship between IQ in siblings had ever used Wechsler's measures. In their review Erlenmeyer-Kimling and Jarvik reported 35 studies of siblings reared together and 2 studies of siblings reared apart. At our recent request, they kindly sent us their bibliographic references for each of these 37 studies plus a summary list of the intelligence tests used by each of the investigators.

Their list revealed that the 37 sibling studies had been reported in 31 different publications between 1912 and 1962 and that a total of 15 different tests had been used in 37 sets of combinations. Additionally, 4 of the 37 studies failed to identify which test(s) yielded the sibling correlation reported. Not surprisingly for this early era, the test most frequently used in studying the correlation between IQs of siblings was the Stanford-Binet (13 studies). This test was followed by the Otis (6 studies), Army Alpha (4 studies), the Terman Group Test (2 studies), and the National Intelligence Test (2 studies). Finally, each of the following tests were used in 1 study: Army Beta, IER Test of Selective and Rational

Thinking, South Africa Group Intelligence Tests, Vocabulary tests from the University of Minnesota College Aptitude Test, Haggerty Intelligence Examination, College Entrance Test, Cattell scales, Pintner Rapid Survey Intelligence Test, Test mosaïque de Gille, and the CVB (Swedish abbreviated Wechsler-Bellevue). It is clear from this that excluding the Stanford-Binet, many of these tests are paper-and-pencil instruments that were designed for self or group rather than individual administration.

The background for the present article is that since 1959 we have been involved in a patrolman selection program for the city of Portland. The 6-hour patrolman examination includes a number of standard clinical psychological assessment instruments, one of which is the individually administered Wechsler Adult Intelligence Scale (WAIS). Noting recently that to date we had examined some 1,200 applicants between the ages of 21 and 31, we researched these records and found that 10 pairs of these applicants were brothers. Although relatively few in number, we decided to compute the correlation between pairs of WAIS scores on these 20 brothers for several reasons. First, because of the value of such data. As Alstrom (1961) has noted:

By comparison with the immense quantity of (intelligence) test investigations which have been carried out all over the world, there have been extremely few carried out on representative series of families. . . . Every such study of a representative sample of families is therefore of great value. (p. 199)

Second, as a test of measured intelligence, the WAIS is among the best such measures currently available. Third, unlike many studies of measured intelligence in families that individually examined (often using student examiners) the two family

Requests for reprints should be sent to Joseph D. Matarazzo, Department of Medical Psychology, University of Oregon Health Sciences Center, Portland, Oregon 97201.

members in one sitting and thus conceivably could introduce a bit of examiner halo effect or other bias, our "study" was neither planned nor were the siblings examined on the same day. That is, except for a single pair of fraternal twins among our 10 pairs who were examined on the same day, the WAIS examiner (ANW) for the present study, a board-certified clinical psychologist, did not know at the time he examined the second applicant that the applicant's brother also had been examined by him on a previous occasion.

After we identified them, the 10 pairs of brothers were divided into Group A (first brother) and Group B (second brother). Descriptive data for the day of examination for Group A and Group B yielded the following means and ranges (in parentheses), respectively: ages = 24.0 years (21-32) and 24.6 years (22-29), and education = 13.2 years (12-16) and 13.7 years (12-17). The interval between testing of Brother A and Brother B ranged from 0 to 83 months, with a mean of 17.4 months and a median of 10.5 months. Thus our study involved 10 pairs of 24-year-old brothers with a mean of 13 years of education who were individually examined with a median interval of 10 months by a highly experienced clinical psychologist who was unaware of the consanguineous relationship.

Followers of Pearson's studies on sibling correlation are aware that in his research he computed a double-entry Pearson correlation, entering each pair of scores twice; first as XY then as YX. His procedure yielded the following values for our data: $r_s = .31, .05$, and $.48$ ($p = .02$), for Full Scale IQ, Verbal IQ, and Performance IQ, respectively.

We were less interested in such statistical sophistication, being more interested here in a standard Pearson correlation to satisfy our clinical curiosity as to what value it would reach for the WAIS given to 10 pairs of brothers. The value of the standard Pearson correlation that we obtained for the Full Scale IQ was $.42$. Although

a bit short of statistical significance with our N of 10 pairs, this is a value not unlike the median correlation of $.49$ yielded in the 37 studies reviewed by Erlenmeyer-Kimling and Jarvik (1963) despite the fact that our 20 subjects earned IQs only in the top 50% of the range (i.e., from a low Full Scale WAIS IQ of 105 to a high of 126, with a mean of 112.6 for Group A and a mean of 115.2 for Group B). The corresponding correlation for Verbal IQ of our 10 pairs was $.12$ and for Performance IQ was $.54$. Other than small sample size, we could find no explanation in our data, such as one or two deviant cases, to explain this lower correlation for Verbal IQ.

The data from the 13 studies that consisted of individually administered Stanford-Binets were probably the most robust among the 37 studies in the review by Erlenmeyer-Kimling and Jarvik (1963). The correlation of $.42$ for Full Scale IQ obtained by us with the equally robust WAIS in these 10 pairs of adult brothers adds one additional datum, suggesting that the median correlation of $.49$ for siblings in the 37 studies reviewed by Erlenmeyer-Kimling and Jarvik is a reliable value. This latter interpretation is especially important inasmuch as our examiner was unaware that he was examining a sibling in all but one pair, and thus our gnawing concern since we began rereading this literature on a potential examiner halo effect in some of the earlier studies was put to rest by the results of the present study.

References

- Alstrom, C. H. A study of inheritance of human intelligence. *Acta Psychiatrica et Neurologica Scandinavica*, 1961, 36, 175-202.
- Erlenmeyer-Kimling, L., & Jarvik, L. F. Genetics and intelligence: A review. *Science*, 1963, 142, 1477-1479.

Received April 11, 1977 ■

Self-regulatory Effects of Monitoring Sensory and Affective Dimensions of Pain

Kenneth D. Craig, Helen Best, and J. Allan Best
University of British Columbia, Vancouver, Canada

The differential effectiveness of focused attention on either affective-discomfort or sensory components of electric-shock-induced pain experiences was examined. In addition, the impact of models displaying tolerance or intolerance for pain was examined on reports of these different components of the experience. Attention to sensory components led to lower pain tolerance, avoidance of shocks at lower current levels, and characterizations of the shocks as more painful than stronger current intensities accepted during the discomfort rating task. The strategy of modeling tolerance was effective in influencing both sensory and affective reports of the experience.

Recent conceptualizations of pain experience distinguish between sensory-discriminative and affective-motivational qualities. Experimental and clinical research indicate that the discrimination between sensory qualities and affective attributes is readily accomplished. This study examined whether a potent form of social influence, exposure to tolerant or intolerant models, would differentially influence characterizations of physical as contrasted with affective qualities of pain experience.

The social context was expected to be a more potent determinant of characterizations of discomfort than physical qualities. This is because subjective discomfort appears to be more inclusive of various experiential components of pain than characterizations of physical qualities, since they are the product of sensory, affective, and contextual factors. In addition, reports of discomfort appear to have more important social functional value and, consequently, should reflect greater sensitivity to the social context.

The investigation also examined the effects of self-monitoring of either discomfort or physical intensity qualities on pain tolerance. Demonstration of differential attentional effects could have some impact on the choice of cognitive strategies to be used for the self-management of pain.

Subjects were 30 male undergraduate student volunteers. They participated in sessions with a

male model portraying a naive subject. He did not receive shocks, and his ratings were based on subjects' responses. Subjects were assigned randomly to one of six groups based on a 3×2 factorial. The first variable concerned whether the model assumed the role of a tolerant or intolerant companion also subjected to shocks or an inactive observer. The second variable was based on the two possible sequences of rating the electric shocks on physical and discomfort intensity judgmental tasks.

Electric shocks of a 1-sec duration were generated by a 60-Hz stimulator and delivered through concentric electrodes to the volar surface of the right forearm. Three ascending series of shocks were delivered with each trial increasing by .5 mA until subjects reported that they could endure no further increases.

Rating scales were described to subjects that provided appropriate labels for characterizing the shocks in terms of either physical intensity or affective discomfort. For judgments of physical intensity, labels read *undetectable*, *low*, *moderate*, *high*, and *extremely high*. For discomfort ratings, labels were *not uncomfortable*, *uncomfortable*, *very uncomfortable*, *painful*, and *intolerable*. It was emphasized that the last switch was to be reserved for a current level beyond which they could endure no further shocks and that no shocks would be given after it was used. All subjects undertook both judgmental tasks in counterbalanced order.

The model enacted the intolerant role by remaining one label ahead of the subject on the intensity scale. The tolerant role was enacted by essentially having the monitor remain one label behind the subject.

Additional questionnaires required evaluations

This investigation was supported by research grants from the Canada Council.

Requests for reprints and for an extended report of this study should be sent to Kenneth D. Craig, Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5.

of the shocks on intensity and discomfort dimensions and ratings concerning other aspects of the study.

Four-way analyses of variance were used to evaluate the effects of modeling, rating order, the two judgmental tasks, and the repeated shock series on pain tolerance. Significant differences were observed as a result of the two judgmental tasks, $F(1, 24) = 4.84$, $p < .05$, and the two modeling roles, $F(2, 24) = 4.75$, $p < .01$. Tukey (*b*) tests ($\alpha = .01$) indicated that the tolerant group accepted substantially greater shocks ($M = 10.41$ mA) than the intolerant group ($M = 6.68$ mA) and the no-model group ($M = 5.97$ mA), whereas these latter two groups did not differ from each other ($p > .10$). There were neither order nor trial effects.

Modeling effects were consistent whether judgments of physical intensity or subjective discomfort were required, with the biasing effects generally consistent across the range of judgmental categories. The tolerant-modeling impact was sufficiently substantial for those exposed to the tolerant model to not characterize as even painful those shocks deemed to be intolerable by subjects paired with an intolerant model or no model.

Tolerance levels were higher when attention was focused on personal discomfort ($M = 8.12$ mA), as contrasted with judgments emphasizing perceived physical intensity ($M = 7.26$ mA). The differential impact of the self-monitoring tasks indicated that the two sets of descriptors referred to discriminable states.

Questionnaire analyses of the severity of discomfort experienced when subjects were involved in either the ratings of physical intensity or affective discomfort indicated that when discomfort was being rated, the most intense shocks were described as less painful than when physical intensity was rated, $F(1, 24) = 7.37$, $p < .05$.

The study clearly demonstrated the substantial

effects of exposure to a tolerant model and that the social influence strategy equally affected characterizations of sensory and affective qualities. Additionally, instructional sets to attend to discomfort led to greater pain tolerance than instructions to describe the sensory quality of the experience, its physical intensity. Lower pain tolerance during intensity ratings was also characterized as more painful than the current intensities that provoked pain tolerance during the discomfort rating task.

A number of possible explanations could account for the superiority of the discomfort-rating task. Attending to personal discomfort may permit a more global and realistic appraisal of the experience. It also may have enhanced perception of the ability to control the experience, because subjective experiences tend to be seen as more amenable to volitional control than sensory experiences. Attending to affective components of experience may also have facilitated preparation for subsequent noxious experience. Restricting attention to the physical intensity scale may have left subjects unprepared for the severity of shocks at more intense levels.

Because clinical pain, in most instances, does not permit deliberate appraisal of low intensity levels, preparatory communications have been used effectively to manage stress reactions. Because of the problems involved in the choice of language to describe most probable reactions to noxious events, the present study indicates the need to instruct individuals to evaluate descriptions of others' experiences within the context of their own responses to noxious stimuli. Attention to different components of the pain experience itself may differentially affect qualities of the experience. Attending to some of the affective qualities of the experience would appear to be an important component of effective self-regulation.

Received May 23, 1977 ■

Comparison of the Similarities and Differences in the Self-concepts of Male Alcoholics and Addicts

Jerome F. X. Carroll, M. Israel Klein, and Yoav Santo
Eagleville Hospital and Rehabilitation Center, Eagleville, Pennsylvania

This study compared self-concept scores from the Tennessee Self Concept Scale (TSCS) of 178 alcohol- and 156 drug-dependent male clients in an abstinent, therapeutic community. A multivariate analysis of the TSCS scores indicated that substance abuse, when age and race were statistically controlled, yielded only three significant results (True/False ratio; Psychosis; and Personality Disorder). These data were interpreted as indicating greater similarity than difference for the alcohol- and drug-dependent men regarding self-concept. The data also indicated the value of using a multivariate rather than a bivariate design in examining the relationships between substance abuse patterns and self-concept.

An increasing number of spokespersons within the substance-abuse field have advocated combined treatment, that is, treating alcoholics and drug addicts together in the same rehabilitation program (Carroll & Malloy, 1977). This proposal, however, raises a number of important questions, one of which is, to what extent are the personality dynamics of these two groups similar/dissimilar?

Even though research studies have shown that alcoholic or drug-dependent persons are low in self-esteem (e.g., Fitts, Arney, & Patton, 1973; Robinson, 1973), there has been little data published comparing the self-concepts of alcoholics and drug-dependent clients seeking treatment. We attempted to fill this gap by comparing the Tennessee Self Concept Scale (TSCS; Fitts, 1965) scores of alcoholic and drug-dependent men while statistically controlling for the effects of age and race.

The present study was undertaken at Eagleville Hospital and Rehabilitation Center, a residential therapeutic community whose abstinence treatment program serves a near-equal number of alcoholics and drug addicts. For a period of 6 consecutive months, all male clients with a diagnosis of either alcoholism or drug dependence who completed the TSCS as part of the standard group test battery at EHRC were included in this study.

Clients with a dual diagnosis of alcoholism and drug dependence were excluded from the study. So too were clients with severe reading problems, which prevented them from completing the TSCS.

No attempt was made to differentiate among alcoholics regarding their preference for beer, wine, or liquor, nor did we attempt to differentiate among the varieties of drug-dependent clients. Most of the drug-dependent clients had been referred for treatment due to heroin abuse, although they often had abused other drugs as well (e.g., marijuana and minor tranquilizers). Classification of clients as being either alcoholic or drug dependent was based solely on the extant system used by Eagleville Hospital and Rehabilitation Center's Medical Records Department for classifying newly admitted clients.

The sample consisted of 334 male clients; 178 had been diagnosed as alcoholics and 156 as addicts. With respect to race, 170 of the men were black and 164 were white. We divided our sample into four age groups (up to 23, 24-33, 34-44, 45 and older) as had been done in a previous study of EHRC clients by Barr, Ottenberg, and Rosen (Note 1).

A contingency analysis of the relationship between age and addiction indicated that these variables were not independent, $\chi^2(3) = 14.91$, $p < .005$. Not surprisingly, younger clients were significantly more likely to be drug addicts, and older clients were more likely to be alcoholics.

Similarly, a contingency analysis of the relationship between race and addiction indicated that these variables also were not independent, $\chi^2(1) = 11.41$, $p < .001$. Black clients were significantly

Requests for reprints and for an extended report of this study should be sent to J. F. X. Carroll, Director Psychological Services, Eagleville Hospital and Rehabilitation Center, Eagleville, Pennsylvania 19408.

more likely to be alcoholics, and whites were more likely to be addicts.

These results differ somewhat from reports received from the National Institute on Drug Abuse (NIDA) and the National Institute of Alcohol Abuse and Alcoholism (NIAAA). Reports from NIAAA indicate more white than black alcoholics are receiving treatment in federally funded programs, although we believe the extent of the alcohol abuse problem among blacks has been underestimated by these data. The distribution of races for drug abuse reported by NIDA correspond with the findings we have reported.

Having demonstrated that addiction was not independent of either age or race in our sample, a three-way analysis of variance was performed for each of the 29 TSCS scales to ascertain the unique contribution of addiction, age, and race, and their interaction effects on TSCS scores.

Addiction, as a main effect independent of race and age, yielded significant main effects for three TSCS scales, namely True/False ratio, $F(1, 319) = 3.51, p < .05$; Psychosis, $F(1, 319) = 4.07, p < .05$; and Personality Disorder, $F(1, 319) = 4.07, p < .05$. Alcoholics scored higher on the T/F ratio, indicating a more acquiescent response set, which we interpret to reflect a more passive and compliant coping style. The higher scores of the alcoholics on the Psychosis and Personality Disorder (empirical) scales indicated a greater degree of emotional distress, poorer reality contact, greater depression and emotional lability, greater mental confusion, higher levels of suspicion, and more personality weaknesses and vulnerabilities than that observed for addicts.

Race as a main effect, independent of substance of abuse and age, yielded two significant main effects, Total Conflict, $F(1, 319) = 5.14, p < .05$, and Personality Integration, $F(1, 319) = 4.00, p < .05$. Black males evidenced greater confusion, contradiction, and general conflict in self-perception than did white males. Black males also scored lower than white males on the Personality Integration scale.

There were no significant main effects due to age. This finding is in sharp contrast to that reported in a companion study (Carroll, Santo, & Klein, Note 2) using the Personality Research Form as a measure of normal personality needs. Apparently age exercises a greater influence when normal personality needs are measured and examined.

Concerning interaction effects, only two sig-

nificant results were observed. A significant interaction for race and age on the Behavior scale was obtained, $F(3, 319) = 2.98, p < .05$, with white males in the 24-33 age group scoring highest, and black males 23 years or younger and 45 years or older scoring lowest. We interpret the significant interaction effect as indicating that the greatest degree of guilt regarding behavior was borne by the two extreme age groups of blacks.

A significant interaction effect for race and addiction was observed on the Personality Integration scale $F(1, 319) = 4.94, p < .05$. In this instance, black alcoholics scored lowest, and white addicts scored highest.

While not minimizing the differences noted above, it is nonetheless important to note that nearly all of the scales that had been significant in various preliminary bivariate analyses (e.g., comparing race, age, or substance abuse and self-concept) ceased to be significant in the multivariate analysis. When race and age are controlled, therefore, alcohol- and drug-dependent men appear to be more similar than dissimilar with respect to their self-concepts.

Reference Notes

1. Barr, H. L., Ottenberg, D. J. & Rosen, A. *The cross-use of alcohol and drugs by addicts and alcoholics: Patterns of abuse of alcohol and drugs in a group of hospitalized drug addicts*. Paper presented at the International Conference on Alcoholism and Addiction, Dublin, Ireland, October 1971.
2. Carroll, J. F. X., Santo, Y. & Klein, M. I. *A comparison of personality characteristics of male alcoholics and addicts as measured by the Personality Research Form*. Unpublished manuscript, Eagleville Hospital and Rehabilitation Center, 1976.

References

- Carroll, J. F. X., & Malloy, T. Combined treatment of alcohol and drug dependent persons: A literature review and evaluation. *American Journal of Drug and Alcohol Abuse*, 1977, 4, 343-364.
- Fitts, W. H. *Tennessee Self Concept Scale manual*. Nashville, Tenn.: Counselor Recordings and Tests, 1965.
- Fitts, W. H., Arney, J., & Patton, W. O. *A self-concept study of alcoholic patients*. Nashville, Tenn.: Dede Wallace Center, 1973.
- Robinson, J. S. *The self-concepts of drug abusers*. Nashville, Tenn.: Dede Wallace Center, 1973.

Received May 26, 1977 ■

Dimensions, Patterns, and Personality Correlates of Drug Abuse in an Offender Population

Terrill R. Holland
California Department of Corrections, Chino

Twelve drug abuse scores obtained from each of 215 prisoners were factor analyzed, resulting in two factors describing the lifetime degree of use of cannabis versus opiate types of drugs. Multiple discriminant analysis of Minnesota Multiphasic Personality Inventory (MMPI) profiles versus drug abuse patterns indicated a moderate, unidimensional relationship between these two sets of variables ($R_c = .38$, $p < .05$). The MMPI profiles of both groups of opiate users were configurally similar to, though less highly elevated than, those identified in previous research with narcotic addicts, suggesting certain differences between the present sample and drug abuse cases encountered in other settings.

Previous research on the average Minnesota Multiphasic Personality Inventory (MMPI) profile and the common profile types among drug-dependent individuals has focused primarily on narcotic addicts and has led to disagreement over questions of both the existence of personality patterns peculiar to addicts and the relative importance of psychopathic versus neurotic and/or schizoid components of addict profiles. However, drug abuse is a multidimensional phenomenon, and the classification of subjects based on multiple measurements reflecting types of drugs and parameters of use would be expected to yield more accurate descriptions of drug abuse patterns and correlates than has been possible by comparing the profiles of extreme groups of individuals designated as either addicts or non-addicts.

In light of the foregoing considerations, the drug abuse histories of 215 prisoners undergoing felony presentence evaluation at the California Institution for Men, Chino, were quantified by scoring each of three major classes of drugs: opiates, nonopiate "hard" drugs, and cannabis)

Grateful acknowledgment is given to Hung Tran for data processing assistance and to Kathleen McKay for clerical support.

Although conducted under the auspices of the Department of Corrections, the opinions expressed here are the views of the author and do not necessarily reflect the official position of the California Department of Corrections or the Health and Welfare Agency.

Requests for reprints should be sent to Terrill R. Holland, California Institution for Men, Box 128, Chino, California 91710.

for four parameters of use (frequency of use, method of administration, age at first exposure, and years of use) using a modification of the scale described by Gunderson, Russell, and Nail (1973). The 12 scores obtained in this fashion were standardized and subjected to a principal components analysis with varimax rotation, resulting in two factors with eigenvalues greater than one. As shown in Table 1, cannabis parameters load most highly on Factor 1, opiate parameters load on Factor 2, and nonopiate hard drug parameters load on both factors, indicating that a two-dimensional model of drug abuse is appropriate for the present subjects and method of measurement. Since parameters of use are not statistically distinct from each other, and since the use of nonopiate hard drugs is not distinct from the use of cannabis and opiates, the 12 drug scores reduce to two underlying dimensions describing the lifetime degree of use of cannabis versus opiate types of drugs.

Calculating factor scores for individuals and dividing the distributions of these scores at the median resulted in four groups defined by levels and combinations of the two drug abuse dimensions. The groups and their associated sample sizes and MMPI Welsh codes are as follows: low cannabis-low opiate ($n = 79$) = $49 - 5268371 / KFL$; high cannabis-low opiate ($n = 50$) = $49 - 8562731 / KFL$; low cannabis-high opiate ($n = 24$) = $428967 - 531 / F - KL$; and high cannabis-high opiate ($n = 62$) = $49578 - 2631 / FKL$. Multiple discriminant analysis of these profiles yielded a single significant dimension of group difference ($R_c = .38$; $\Lambda = .78$), $\chi^2(36) = 51.31$, $p < .05$, the composi-

Table 1

Varimax Rotated Factor Loadings for Drug Abuse Scores

Score	Factor 1	Factor 2
Opiate rate	.241	.934
Opiate method	.250	.935
Opiate age	.241	.889
Opiate duration	.195	.919
Hard drug rate	.680	.594
Hard drug method	.664	.641
Hard drug age	.718	.513
Hard drug duration	.659	.600
Cannabis rate	.875	.200
Cannabis method	.905	.158
Cannabis age	.864	.156
Cannabis duration	.883	.280
Eigenvalue	8.208	1.910
Proportion of variance	.684	.159

tion of which may be interpreted on the basis of "structure coefficients," that is, bivariate correlations between the total discriminant score and each of the original variables (Cooley & Lohnes, 1971). In this respect, the groups are maximally discriminated along a dimension that is correlated .30 or greater with Scales $F(r = -.39)$, $D(r = -.32)$, $Pd(r = -.73)$, $Pa(r = -.30)$, $Pt(r = -.48)$, and $Sc(r = -.36)$, with selective opiate users manifesting the highest score in each instance. One-way univariate analyses of variance resulted in statistically significant F ratios for D ($p < .05$), Pd ($p < .001$), and Pt ($p < .05$).

The findings indicate a moderate, unidimensional relationship between pattern of drug abuse and MMPI performance. To the extent that these two sets of variables are related to each other, nondrug users and selective cannabis users exhibit the least pronounced personality liabilities, whereas selective and nonselective opiate users manifest the most conspicuous deficiencies. Further, although all four group profiles are suggestive, to a greater or lesser degree, of social nonconformity, the selective opiate group is characterized by mild, though noteworthy, additional elements of subjective distress and disturbed thinking.

The configurations of the MMPI profiles of

both of the present groups of opiate users are similar to the major profile types identified through cluster analysis by Berzins, Ross, English, and Haley (1974) among narcotic addicts, thus providing additional support for the hypothesis that opiate dependence occurs within more than one personality context. Nonetheless, the present profile elevations are considerably lower than those obtained in both the Berzins et al. study and in most previous research on the relationship between drug abuse and MMPI performance. These differences might be attributed partially to the profile-suppressing variables of detoxification, nonvolunteer status, and incarceration in a relatively protected environment. However, this trend is also consistent with the tendency of presentence cases of the type studied to exhibit a generally lesser degree of psychopathology on the MMPI than most other groups of incarcerated offenders (Holland & Holt, 1975). Factors probably responsible for this phenomenon are the motivation of the subjects to present themselves in a favorable light to their evaluators, optimism resulting from the possibility of returning home on probation in the foreseeable future, and an element of selectivity in which many offenders who are clearly emotionally disturbed are often incorrectly perceived by the courts as being dangerous and therefore are given lengthy sentences without first being referred for presentence evaluation through the state prison system.

References

- Berzins, J. I., Ross, W. F., English, G. E., & Haley, J. V. Subgroups among opiate addicts: A typological investigation. *Journal of Abnormal Psychology*, 1974, 83, 65-73.
- Cooley, W. W., & Lohnes, P. R. *Multivariate data analysis*. New York: Wiley, 1971.
- Gunderson, E. E., Russell, J. W., & Nail, R. L. A drug involvement scale for classification of drug abusers. *Journal of Community Psychology*, 1973, 1, 339-403.
- Holland, T. R., & Holt, N. Personality patterns among short-term prisoners undergoing presentence evaluations. *Psychological Reports*, 1975, 37, 827-836.

Received June 20, 1977

MMPI Evaluation of 5-Year Methadone Treatment Status

Gennaro Ottomanelli, Peter Wilson, and Richard Whyte
Department of Psychiatry, Downstate Medical Center
State University of New York, Brooklyn

Treatment outcome over a 5-year period was evaluated for 148 first admissions to a methadone treatment program. Eleven patients (7%) were successful treatment completions, 16 patients (11%) transferred to other methadone programs, 38 patients (26%) remained in continuous treatment, and 83 patients (56%) were unsuccessful treatment terminations. Discriminant analysis using the Minnesota Multiphasic Personality Inventory (MMPI) suggested that the more stable patients at admission had the best treatment outcome. For the patients in continuous treatment, MMPIs administered at 6-week, 6-month, and 5-year intervals indicated that this group of patients did not change on the personality dimension.

The questions formulated for research were (a) Is the Minnesota Multiphasic Personality Inventory (MMPI) useful in predicting long-term treatment outcome for methadone patients? Also, does the MMPI present different group characteristics representative of different treatment outcomes? and (b) Given a group of methadone patients in long-term treatment, does the MMPI demonstrate change over time on the personality dimension?

The sample consisted of 148 narcotic addicts admitted to a methadone clinic at the Kings County Addictive Disease Hospital, Brooklyn, New York, during 1971. The sample represented voluntary first admissions to the methadone clinic. The criteria for program admission were 2 or more years of admitted drug use, age over 20 years, and absence of overt psychosis (determined by psychiatric evaluations). The sample consisted of 110 (74%) males and 38 (26%) females with a mean age of 27.14 years ($SD = 6.26$); 73 (49%) were white, 43 (29%) were black, and 32 (22%) were Hispanic. The sample had a mean of 2.19 months ($SD = 3.51$) of employment in the 1 year prior to admission, a mean of 3.11 arrests ($SD = 4.02$) prior to admission, and a mean length of drug use of 94.07 months ($SD = 55.8$ months).

The MMPI and a questionnaire eliciting demographic information were administered

within 6 weeks after admission. Another questionnaire eliciting employment and arrest information, housing stability, current sources of income, and methadone-related attitudes supplemented the MMPI administration at the 5-year follow-up. MMPI profiles were classified as invalid on the basis of an F score ≥ 22 .

Treatment outcome over a 5-year period for the group of 148 patients showed that 11 patients (7%) completed treatment, 38 patients (26%) were in continuous treatment for 5 years, 83 patients (56%) were unsuccessful treatment terminations (patients discharged for an assortment of administrative and disciplinary reasons), and 16 patients (11%) had transferred to other methadone programs. Of the 148 patients admitted to the study, 38 patients did not cooperate with the MMPI testing. This patient group consisted of 29 males (76%) and 9 females (24%); 14 patients (37%) were white, 15 (39%) were black, and 9 (24%) were Hispanic. Five-year treatment outcome for this group showed that 3 patients (8%) were treatment completions, 4 patients (10%) were in continuous treatment for 5 years, 29 patients (76%) were unsuccessful treatment terminations, and 2 patients (5%) transferred to other programs.

The remaining 110 patients completed the MMPI within 6 weeks after admission. Five-year treatment outcome for this group showed that 8 patients (7%) were treatment completions, 34 patients (31%) were in continuous treatment for 5 years, 54 patients (49%) were unsuccessful treatment terminations, and 14 patients (13%) transferred to other programs. The demographics of the 110 MMPI patients

Requests for reprints and for an extended report of this study should be sent to Gennaro Ottomanelli, Department of Psychiatry, Downstate Medical Center, State University of New York, Box 9, Code 24, 450 Clarkson Avenue, Brooklyn, New York 11203.

classified according to treatment outcome are as follows: (a) The 8 patients in the treatment-completed group consisted of 7 males (88%) and 1 female (12%); 3 were white (37%), 3 were black (37%), and 2 were Hispanic (25%). (b) The in-treatment group consisted of 25 males (74%) and 9 females (26%); 20 were white (59%), 5 were black (15%), and 9 were Hispanic (26%). (c) The unsuccessful treatment group consisted of 41 males (76%) and 13 females (24%); 27 were white (50%), 15 were black (28%), and 12 were Hispanic (22%). (d) The transferred group contained 14 patients, 7 males (50%) and 7 females (50%); 9 were white (64%) and 5 were black (36%).

Discriminant analysis was used to evaluate treatment outcome based on the MMPIs administered in the 6th week after admission. The groups were sorted on the basis of 5-year treatment outcome, that is, treatment completed (successful treatment effort), currently in treatment, and unsuccessful treatment terminations. One patient, in each of the treatment completed, active, and unsuccessful treatment groups was eliminated from the discriminant analysis because of invalid MMPIs ($F \geq 22$). A chi-square analysis compared the accuracy of the classification matrix of the discriminant functions with the expected frequencies based on chance and was found to be statistically significant, $\chi^2(4) = 27.83$, $p < .01$.

Comparison of the means of the MMPI scales for the patients grouped on the basis of treatment outcome showed that the patients in the successful treatment group ($n = 7$) had their highest mean T scores on the Pd (70.00) and Ma (66.29) scales, whereas the remaining eight MMPI clinical scales ranged from a mean of 64 (D) to a mean of 50.14 (Si). For the patients who remained in treatment ($n = 33$), the highest mean T scores were obtained on the D (71.48) and Pd (74.21) scales, and the remaining eight MMPI clinical scales ranged from a mean of 66.03 (Ma) to a mean of 57.36 (Si). For the patients in the unsuccessful treatment group ($n = 53$), the highest mean T scores were obtained on the Pd (72.34) and Ma (70.85) scales, and the remaining eight MMPI clinical scales ranged from a mean of 68.07 (Sc) to 54.30 (Si). Generally, the scales satisfied clinical expectations on two counts: (a) The patients with the best treatment outcome, that is, the treatment completed group, were the most stable at the time of admission, since this group had the lowest mean T score on 8 of 10 clinical scales of the MMPI; and (b) there was a linear trend (i.e., the rela-

tionship approaches a straight line) for the three groups on the F , Pa , Pt , and Sc scales that rank patients on emotional traits, leading to a greater probability of acting out and concomitant greater probability of unsuccessful treatment.

Although attempts were made to retest the 34-patient MMPI in-treatment group, only 24 patients in this group provided valid 6-week and 5-year MMPIs; within this subgroup were 13 patients who completed the MMPI at 6-week, 6-month, and 5-year intervals. Although the 24-patient group showed increased scores on K , Hs , D , Hy , Pd , Mf , Pa , Pt , and Sc from the 6-week to the 5-year testing interval, Hotelling's T^2 for a one-sample test-retest design using 12 scales (K was excluded from the analysis) was not statistically significant, $F(12, 12) = 2.06$, $p > .05$.

Since the 13-patient subgroup had been tested on three occasions in the follow-up period, an additional analysis was conducted on these three test occasions. The comparison of 12 MMPI scores (the K scale was excluded) for 6 weeks and 6 months was not statistically significant; Hotelling's T^2 for one sample, $F(12, 1) = 47.82$. Likewise the comparison of the 12 MMPI scores for the 6-month and 5-year intervals was not statistically significant; Hotelling's T^2 for one sample, $F(12, 1) = 2.24$.

Aside from the patients who transferred to other institutions, the MMPIs of the three groups appear to be in accord with clinical expectations; that is, the most stable group at admission consisted of those patients who successfully completed treatment. Those patients who remained in treatment had their highest MMPI elevations on the D and Pd scales. For this group the pharmacology of methadone as a potent analgesic may have been an influential factor. The unsuccessful treatment group had their highest MMPI elevations on the Pd , Sc , and Ma scales, with the accompanying higher probability of acting out and higher probability of unsuccessful treatment termination.

The follow-up of the 24-patient group in continuous treatment for the 5-year period did not demonstrate substantive gains on the personality dimensions measured by the MMPI. When one considers the analysis of the MMPIs of the 24-patient group and the analysis of the 13-patient subgroup, it appears that the patients remained unchanged on the personality dimensions at 6-week, 6-month, and 5-year intervals. Although the results of the present 148-patient study are subject to the limitations imposed by a one-

sample design with patients cooperating on a voluntary basis, the findings of the present study suggest that methadone stabilization is most effective for the small proportion of patients with personality and social strengths. Methadone patients who are treatment manageable but lacking in personality and social strengths remain in treatment to preserve marginal social function-

ing, with minimal improvement resulting from long-term stabilization. The marginal social functioning of this group could deteriorate if treatment efforts are terminated. The largest group of patients studied were treatment unmanageable and unsuccessful treatment terminations.

Received August 1, 1977 ■

Stimulus-Seeking Behavior in Three Delinquent Personality Types

David A. Shostak
University of Virginia

Curtis W. McIntyre
Southern Methodist University

This study examined stimulus-seeking behavior in three delinquent personality types (psychopathic, neurotic, and socialized) drawn from three populations (juvenile delinquent, young adult offenders, and college). Results from a kinesthetic aftereffect task and the Sensation Seeking Scale indicate only limited evidence of pathological stimulus seeking by the psychopathic delinquent personality type. We suggest that this limited evidence results from the incarceration experiences of the young adult offenders.

Three delinquent personality types (unsocialized psychopathic, disturbed neurotic, and socialized delinquent) have been described by Quay (1965). Moreover, Quay has suggested that the behavior of the unsocialized psychopath results from pathological stimulus seeking.

More recently, individual differences in stimulus-seeking behavior have been investigated by two personality researchers, Petrie (1967) and Zuckerman (1971). Petrie has shown that individuals differ in their tendency to reduce or to augment stimulation received during a kinesthetic aftereffect task (KAE). Reducers subjectively diminish their sensory input and seek higher levels of stimulation. Augmenters subjectively increase their sensory input and avoid higher levels of stimulation. Zuckerman has shown that individuals differ in the optimal level of stimulation they seek as measured by his Sensation Seeking Scale (SSS).

In the present study, Quay's suggestion that the unsocialized psychopath is a pathological stimulus seeker was tested by administering the KAE and SSS to all three delinquent personality types. The general expectation was that unsocialized psychopaths would be reducers on the KAE and high scorers on the SSS.

Three male populations were used: juvenile offenders (M age = 15 years 2 months), young adult offenders (M age = 20 years 6 months), and college students (M age = 19 years). Ten subjects of each delinquent personality type were selected within each population using the Quay and Peterson Personal Opinion Questionnaire. Following selection, each subject took the KAE and SSS.

Requests for reprints and for an extended report of this study should be sent to Curtis W. McIntyre, Department of Psychology, Southern Methodist University, Dallas, Texas 75275.

Separate one-way analyses of variance were applied to each population to assess stimulus-seeking differences between the three delinquent personality types. With respect to the KAE, the only significant difference was found for the young adult offenders, $F(2, 36) = 3.39, p < .05$. On this task the unsocialized psychopaths reduced most; that is, they sought higher levels of sensory input. A general tendency to reduce was observed for both the juvenile and young adult offender populations.

With respect to the SSS, no differences between the three delinquent personality types were found for any of the three populations tested. Evidently, sensation seeking (as measured by the SSS) was uniform across all three diagnostic categories. In addition, correlations between the KAE and each subscale of the SSS failed to reach significance for any of the personality types and populations. Evidently, the stimulus-seeking behaviors assessed by these measures were independent.

In conclusion, only limited evidence of pathological stimulus seeking by Quay's psychopathic delinquent personality type was found. Moreover, we suggest that this limited evidence may not be etiological in nature but may result from the longer incarceration (or isolation) experiences of the young adult offenders.

References

- Petrie, A. *Individuality in pain and suffering*. Chicago: University of Chicago Press, 1967.
- Quay, H. C. Psychopathic personality as pathological stimulus seeking. *American Journal of Psychiatry*, 1965, 122, 180-183.
- Zuckerman, M. Dimension of sensation seeking. *Journal of Consulting and Clinical Psychology*, 1971, 36, 45-52.

Received August 9, 1977 ■

Validity Generalization of the WISC-R Factor Structure with 10½-Year-Old Children

David A. Shiek and John E. Miller
Western Kentucky University

The purpose of this study was to investigate the robustness of the Wechsler Intelligence Scale for Children-Revised (WISC-R) factor structure. The generalization sample was composed of 126 10½-year-old children of which 62 were male and 64 were female. A principal-components method of factor analysis yielded three reliable factors. Comparisons of the loadings obtained with the generalization sample and the 10½-year-old national standardization sample suggest a high degree of similarity in composition, magnitude, and pattern. The findings highly support the robustness of the WISC-R's factor structure across divergent 10½-year-old samples.

The purpose of this study was to investigate the robustness of the factor structure of the Wechsler Intelligence Scale for Children-Revised (WISC-R). The factor analysis of the standardization data (Kaufman, 1975) yielded three stable factors: Verbal Comprehension, Perceptual Organization, and Freedom from Distractibility. The present sample was composed of 126 children with a mean age of 10.6 years from lower to lower-middle-class homes in the central south-eastern United States. The sample was composed of 62 males and 64 females of which 87 were white and 39 were black. A preliminary analysis indicated two basic differences between this sample and the standardization sample: The Verbal, Performance, and Full Scale IQs were significantly lower, and the variances on the Performance and Full Scale variables were significantly restricted.

The primary analysis consisted of a principal-components method of factor analysis with squared multiple correlations in the diagonals and a varimax rotation procedure. This procedure yielded three reliable factors with eigenvalues greater than 1.0. The first factor (Verbal Comprehension) consisted of Information, Vocabulary, Similarities, and Comprehension. The second factor (Perceptual Organization) was composed of Block Design, Object Assembly, Picture Completion, Mazes, and Picture Arrangement. The

third factor (Freedom from Distractibility) included Coding, Arithmetic, and Digit Span. Comparisons of the factor structure obtained with the generalization sample (this study) and Kaufman's (1975) 10½-year-old national standardization group were made. Visual comparisons of the loadings obtained on the two analyses suggested a high degree of similarity in the loadings. Vector comparisons yielded an intraclass coefficient of .89, indicating a high degree of similarity in pattern and magnitude in the factor loadings. Matrix comparisons indicated that the three factors were similarly delineated in both analyses.

When compared to Kaufman's (1975) findings and conclusions, the factor structure obtained with the generalization sample was highly consistent. Only insignificant variations in the two solutions were present, and the factor structure of the WISC-R appeared quite stable across the two relatively divergent 10½-year-old samples. The conclusions highly supported the robustness of the WISC-R's factor structure and Wechsler's original proposition that his scales assess verbal, performance, and nonintellectual functions. The existence of such a stable factor structure suggests the possibility of exploration as to the clinical usefulness of factor scores in differential diagnosis and prediction.

Reference

- Kaufman, A. S. Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology*, 1975, 43, 135-147.

Received September 14, 1977 ■

Requests for reprints and for an extended report of this study should be sent to David A. Shiek, Department of Psychology, Western Kentucky University, Bowling Green, Kentucky 42101.

Comparison of Self-administered and Examiner-administered Depression Adjective Check Lists

Bernard Lubin
University of Missouri at Kansas City

Joseph G. Marone
Norwich Hospital
Norwich, Connecticut

Ronald G. Nathan
University of Houston and
Texas Research Institute of Mental Sciences, Houston

Self-administered and examiner-administered Depression Adjective Check Lists were compared. One half of each equivalent form A, B, C, D was administered in the standard manner and one half was read by the examiner to 64 male and 64 female psychiatric inpatients. The Wechsler Adult Intelligence Scale Vocabulary subtest was administered at the same time. A repeated measures analysis of variance revealed no significant effects for vocabulary (median split), sex, or method of administration. The significant main effect of lists seems best understood as a chance finding. The results support the use of the examiner-administered method in cases of functional illiteracy.

The standard method of administration of a number of self-administered personality measures used in research and in clinical practice is not feasible in the case of persons who are functionally illiterate. The possible effects of substituting examiner administration for self-administration is an important but relatively unexplored question. Earlier work comparing the two methods of administering the Rotter Incomplete Sentences Blank (Flynn, 1974) and the Minnesota Multiphasic Personality Inventory (Reese, Webb, & Foulks, 1968) found no significant differences. Unfortunately, these instruments do not have alternate forms, and the findings might be the result of practice effect. In addition, these findings were established with "trait" measures; they might not generalize to "state" measures.

The Depression Adjective Check Lists (DACL; Lubin, 1967) were designed as brief, self-administered measures of transient depressive mood. Their alternate forms (Lubin, 1967; Lubin, Dupre, & Lubin, 1967) make them useful in comparing administration methods.

Sixty-four male and 64 female psychiatric inpatients of a large state hospital were individually

administered Set 1 of the DACL (Forms A, B, C, and D) in random order. Each form was divided so that one column was administered verbally by the experimenter and one column was self-administered (written) by the subject. The order of columns was counterbalanced. The Vocabulary subtest of the Wechsler Adult Intelligence Scale was then administered to study the effect of verbal fluency on DACL scores. A median split of the scaled scores was used to analyze the effect of this factor.

A repeated measures analysis of variance revealed a significant main effect of lists, $F(3, 372) = 23.09$, $p < .001$, but no significant effects of vocabulary, sex, or method of administration on the depression scores. The significant main effect of lists was not consistent with earlier findings (Lubin, 1967; Lubin et al., 1967). It seems best understood as a chance finding and thus uninterpretable.

There was no significant difference between self-administered ($M = 5.80$, $SD = 4.23$) and examiner-administered ($M = 5.53$, $SD = 4.20$) depression scores. In fact, the self-administered scores averaged only about $\frac{1}{2}$ point higher on the standard DACL (range = 0-34) than the examiner-administered score. The trend of $F(1, 124) = 3.42$, $p < .07$, should be interpreted within the perspective of the high intercorrelations among lists across administrations (mean $r = .85$, $n = 12$).

These results strongly suggest that the two methods of administering the DACL are similar,

We wish to acknowledge the assistance of Bob Paver, Jim Millham, Linda Peterson, and Daniel Sheer.

Requests for reprints should be sent to Bernard Lubin, Department of Psychology, 5319 Holmes Street, University of Missouri, Kansas City, Missouri 64110.

and they support the use of the examiner method of administration in cases of functional illiteracy.¹

References

Flynn, W. Oral versus written administration of the Incomplete Sentences Blank. *Newsletter for Re-*

search in Mental Health and Behavioral Sciences, 1974, 16, 19-20.

Lubin, B. *Depression Adjective Check Lists: Manual*. San Diego, Calif.: Educational and Industrial Testing, 1967.

Lubin, B., Dupre, V. A., & Lubin, A. W. Comparability and sensitivity of Set 2 (Lists E, F, and G) of the Depression Adjective Check Lists. *Psychological Reports*, 1967, 20, 756-758.

Reese, P. M., Webb, J. T., & Foulks, J. D. A comparison of oral and booklet forms of the MMPI for psychiatric patients. *Journal of Clinical Psychology*, 1968, 24, 436-437.

¹ These findings might have some implications for other handicaps preventing administration of the Depression Adjective Check Lists in the standard manner, for example, blindness or manual handicap, but it would be necessary in these cases to study the examiner-administered method with these subjects directly.

Received September 27, 1977 ■

Early Childhood Autism and Structural Therapy: Outcome After 3 Years

Alan J. Ward

Henry Horner Children's Center, Chicago, Illinois

The effect of 3 years of structural therapy on 21 inpatient cases of early childhood autism (ECA) is examined. Treatment resulted in the discharge of 12 patients. Details of treatment procedure, therapeutic progress, and their effects on diagnostic and prognostic conceptualizations are presented. Comparisons are made among previous reports of attempted treatment of ECA, as well as the results of two other treatment units in the same setting. Results support the hypothesis that the high stimulation, physically intrusive, gamelike, novelty filled, and developmentally oriented treatment approach of structural therapy is capable of producing a significant improvement in cases of ECA.

This article is a preliminary report on the outcome of the application of structural therapy to the residential treatment of early childhood autism. The term *early childhood autism* (ECA) is used to include both the classically defined rare cases of early infantile autism (EIA), as well as the much more common cases of organic autism and the variously handicapped and/or disturbed children who are mistakenly labeled as suffering from EIA.

The treatment of EIA is a topic that has usually aroused great feelings of futility.

Examination of outcome data reported by both Bettelheim (1967) and by Eisenberg (1956) revealed strongly contrasting findings. Eisenberg's follow-up evaluation was divided into the three categories of poor, fair, or good outcome.

Bettelheim has reported outcome figures on a group of 40 "autistic" children and has used the categories devised by Eisenberg. A good outcome was reported for 17 children (42%); a fair outcome, for 15 children (38%); and a poor outcome, for only 8 patients (20%). These outcome figures are in marked contrast to those reported by Eisenberg of 5% for a good outcome, 22% for a fair outcome, and 73% for a poor outcome.

A new therapeutic approach labeled *structural therapy* (Ward, 1970) was used in the development of a treatment program for EIA children.

Evaluation of the 21 original children in this

program revealed only 4 children who met the research definition of EIA, which was (a) lack of the development of object relations from birth; (b) lack of the use of speech for communication; (c) maintenance of sameness via stereotypic behavior with a rage or withdrawal reaction on interruption, and (d) no major neurological dysfunction. The behavioral characteristic of "lack of affective response" is one that was evaluated and agreed on by both me and the co-director of the program, H. Allen Handford. This evaluation was based on clinical interview, play observation, family interview, and review of clinical referral material that included social history, psychological evaluation, psychiatric evaluation, and pediatric neurological evaluation. The other children fell into the diagnostic categories of childhood schizophrenia (5) primary retardation (7), secondary retardation (3), and developmental retardation associated with diffuse brain damage (2). However, all of these children were found to display the behavioral characteristics of a "lack of affective response," whereas children from all five of the above diagnostic categories were found to display the characteristics of "lack of object relationships," "lack of the use of speech for communication," and of having come from an "unstimulating mother/infant relationship." The 4 EIA cases were distinguished from the other disturbed children by their combination of (a) lack of neurological dysfunction and their (b) maintenance of sameness via stereotypic behavior.

The total research unit was organized according to the precepts of structural therapy. The milieu emphasized spontaneous physical and verbal stimulation applied to the children in a playful and gamelike fashion. The goal of this approach

Requests for reprints and for an extended report of this study should be sent to Alan J. Ward, Henry Horner Children's Center, 6500 West Irving Park Road, Chicago, Illinois 60634.

was to increase the amount of varied and novel stimulation received by these children and to use this increased stimulation to make them more aware of their external environment and to help them to progress from their positions of early developmental fixation. The physical stimulation was used to develop body image and bodily awareness and to help provide the body ego, which appears to be necessary for the development of higher ego functions. Twelve of the 21 children were seen in individual therapy on a twice weekly basis, although often the children were seen on an informal basis five times a week. Selected children were seen for sessions in speech therapy on a twice weekly basis for periods ranging from 3 months to 2 years.

Families were contacted on a weekly basis by the social workers. Group counseling was provided on a biweekly basis for all of the parents, and the majority of them had a weekly day visit, overnight visit, or weekend visit with their child.

The basic goal of this structural therapy treatment program has not been "cure" but development. EIA and ECA are viewed as severe developmental disorders of the same kind as are often observed with the rubella child, the blind child, or the deaf child. The basic etiology is considered to be rooted in a deficit of novel and varied stimulation for multiple reasons.

This treatment program attempted to help these children to progress to the point where they achieved the goals of (a) relationships with people; (b) self-care such as toilet training, feeding, and dressing; (c) communication of simple needs in a consistent manner; and (d) the capacity to follow simple directions. The achievement of these goals revealed a child who was still functioning below age-appropriate level in regard to cognitive and affective behavior. The child was now at a point in development in which more conventional play therapy, speech therapy, and/or special education could be used. It was felt that these children should be viewed as being in the midst of their therapeutic course at the time of discharge. They were discharged at this point because the institution was no longer able to provide the needed therapeutic level of stimulation. Those children who were discharged to their homes in the community seemed appropriate for placement in Eisenberg's category of fair outcome. Each child was used as his or her own control in this research, but some comparisons will also be drawn with two other treatment units, in the same setting, that have attempted to work with autistic children (ECA).

The application of 3 years of structural therapy

to the original group of 21 cases of ECA resulted in the discharge to homes in the community of 12 of the original 21 children (57%). These children were placed in normal nursery schools, special classes for the retarded or the emotionally disturbed in public schools and private schools, and sheltered workshops run by the local association for retarded children. The families were referred to the appropriate agencies for continued counseling.

The experimental unit was Unit A, which had a population of boys and girls with a mean age of 8.9 years and 10.2 years, respectively, at admission and whose mean length of prior hospitalization was 1.08 years and 1.92 years. Inspection of comparable data on Comparison Units B and C revealed little difference as to mean age but a great difference as to length of prior hospitalization. These data suggest that the children on Unit A were more severely disturbed than those on Unit B or Unit C. The boys on Unit B had a mean age of 7.2 years on admission and a mean length of prior hospitalization of 6.5 months, whereas the girls on Unit C had a mean age of 8.9 years on admission and a mean length of prior hospitalization of 5.9 months.

Between September 1966 and September 1969, Unit A discharged 12 cases of ECA to home in the community, whereas Unit B discharged 2 boys and Unit C discharged 5 girls. A comparison of the outcome figures of Eisenberg and Bettelheim, vis-à-vis Units A, B, and C, revealed that Unit A exceeded both Eisenberg's and Bettelheim's results in regard to the percentage of children who achieved a fair outcome. Unit C's outcome rate of 25% approximated the outcome rate reported by Eisenberg (22%), but Unit B's outcome figure of 9% fell markedly below that. A classical, psychodynamically oriented psychotherapy and play therapy treatment was used in both Units B and C.

The outcome figures of this preliminary study appear to support the hypothesis that structural therapy is capable of producing significant therapeutic change in children classified as having ECA.

References

- Bettelheim, B. *The empty fortress*. New York: Free Press, 1967.
- Eisenberg, L. The autistic child in adolescence. *American Journal of Psychiatry*, 1956, 112, 607-612.
- Ward, A. J. The application of structural therapy to the residential treatment of early infantile autism. *Schizophrenia*, 1970, 2, 92-102.

Received October 3, 1977 ■

Coital Position and Sex Roles: Responses to Cross-sex Behavior in Bed

Elizabeth Rice Allgeier
State University of New York, Fredonia

Arthur F. Fogel
Eastern Michigan University

The impact of changes in sex role norms regarding heterosexual interaction was explored by varying both the coital position used by a couple in stimulus slides and the extent to which observers identified with stereotypic sex role norms. Females were more negative toward the couple having intercourse in the woman-above position than they were toward the couple in the woman-below position. Observers' degree of sex typing was unrelated to their reactions to the woman-above couple, suggesting that gender may still be more important than sex typing in determining responses to roles in the context of heterosexual interaction.

Some physicians have claimed that female liberation (and the increased demand for male performance) is the primary cause of an increase in male impotence (Liddick, 1972). On the basis of interviews with 50 men, however, Bry (1975) concluded that men prefer sexually aggressive women. Hunt (1974) suggested that the norm of male dominance and female submissiveness during intercourse has been changing, but to date, we have no experimental evidence to indicate the average person's reaction to female assertiveness in bed, nor do we know attitudes toward the male when he assumes the more submissive position.

Thus, one purpose of the present research was to explore male and female reactions to variations in coital positions relevant to sex role norms. Aside from giving birth, there is probably no other arena in which traditional gender role differences have more importance to us than in our sexual interaction. Thus, it was hypothesized that subjects would make more negative attributions about a couple engaging in women-above intercourse than about the same couple engaging in women-below intercourse.

This research was part of a doctoral dissertation submitted by the first author to Purdue University and was supported by National Science Foundation Grant SOC 74-15254 to Donn Byrne.

The authors gratefully acknowledge the helpful comments of Rick Allgeier, Donn Byrne, Rick Kimball, Don Lehr, Dave Przybyla, and Winnie Shepard on an earlier draft.

Requests for reprints and for an extended report of this study should be sent to Elizabeth Rice Allgeier, Department of Psychology, State University of New York, Fredonia, New York 14036.

The second purpose of the research was to see if observers' degree of sex typing would influence their responses to traditional versus nontraditional coital positions. Bem (1976) has demonstrated that androgynous persons (those who give equally high endorsement to masculine and feminine traits as self-descriptive) are more willing to engage in cross-sex behavior, and are more comfortable while doing so, than are sex-typed persons. Accordingly, it was hypothesized that sex-typed persons would respond more negatively to the couple in the women-above position than would androgynous persons.

Subjects were 119 unmarried, introductory psychology students at Eastern Michigan University who volunteered to participate in a study entitled *Responses to Erotica*. After students had completed a demographic questionnaire and the Bem (1974) Sex-Role Inventory, six slides of a couple engaged in intercourse were projected onto a large movie screen for 1 minute each. All students saw three "neutral" slides of the nude couple lying side by side mutually engaged in foreplay. In addition, half of the students saw three slides of the couple having intercourse in the woman-above coital position, and the other half of the students saw three slides of the couple in the woman-below coital position. Students were then instructed to complete a person perception task by checking a scale position along each of nine 7-point bipolar scales that best reflected their impressions of the woman (man) in the slides with respect to the dimensions of adjustment, cleanliness, respectability, morality, femininity (masculinity), goodness, sophistication, desirability as a wife (husband), and desirability as a mother (father).

Results of the $2 \times 2 \times 2$ (Coital Position \times Students' Gender \times Sex Typing) unweighted means analyses of variance indicated consistent interactions between subjects' gender and the coital position on ratings of the woman's cleanliness, $F(1, 111) = 5.68$, $p < .01$; respectability, $F(1, 111) = 7.96$, $p < .01$; morality, $F(1, 111) = 8.79$, $p < .005$; goodness, $F(1, 111) = 7.45$, $p < .01$; desirability as a wife, $F(1, 111) = 5.39$, $p < .05$; and desirability as a mother, $F(1, 111) = 5.92$, $p < .05$. Gender \times Coital Position interactions were also obtained on ratings of the man's cleanliness, $F(1, 111) = 5.61$, $p < .05$; respectability, $F(1, 111) = 5.74$, $p < .05$; morality, $F(1, 111) = 6.80$, $p < .01$; and masculinity, $F(1, 111) = 7.10$, $p < .01$.

Internal comparisons revealed that females consistently rated the couple in the woman-above position significantly more negatively than they did the couple in the woman-below position. Specifically, females rated the woman as dirtier, less respectable, less moral, less good, less desirable as a wife, and less desirable as a mother when she was on top than when she was beneath the man during intercourse.

Females also rated the man as dirtier, less respectable, less moral, and less masculine when he was in the woman-above position than when he was in the woman-below position. Males, on the other hand, tended to respond more positively to the couple in the woman-above position than in the woman-below position; however, the differences were not significant. Thus, the hypothesis regarding responses to variations in coital position was supported with respect to the responses of females and rejected with respect to the responses of males. No differences emerged as a function of the extent of students' sex typing, so the second hypothesis was rejected.

The finding that females (but not males) discriminated against the woman-above couple is particularly interesting in light of Masters and Johnson's (1966) finding that the female coital response develops more rapidly and with greater intensity in the woman-above position than in any other. The design of the present study does not allow us to determine why females responded as they did to the woman-above couple, and further research should be conducted to determine the extent to which their bias is a result of the belief that males find female assertiveness in bed threatening or unattractive. If this were the case, and if further research replicates the finding that males do not discriminate in their evaluations of a couple based on coital position, such informa-

tion might be very useful to the therapist working with nonorgasmic women who inhibit their own initiative due to fear that they will displease or threaten their mates.

The failure of students' sex typing to interact with coital position is in apparent contradiction to the findings of Bem (1976) that the extent of sex typing influences such diverse behaviors as nurturance, independence of judgment, and willingness to engage in cross-sex behavior. On the other hand, none of her designs have studied the context of heterosexual activity, and it may be that the influence of variations in sex typing does not yet extend to relations between the sexes. Support for this possibility is provided by Zeldow (1976), who found no relationship between sex typing and scores on the Attitudes Toward Women Scale (Spence & Helmreich, 1972). A number of the items in this scale deal with sexual relations, and this may be the area in which women feel most reluctant to abandon the feminine norms with which they were raised. Further research should be aimed at determining if sex typing does influence attitudes and behavior during heterosexual interaction in less explicitly sexual contexts (e.g., initiation of dates, sharing cost of dates, etc.) or if the effect of sex typing is negligible in any context involving male-female interaction.

References

- Bem, S. L. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1974, 42, 155-162.
- Bem, S. L. Probing the promise of androgyny. In A. G. Kaplan & J. P. Bean (Eds.), *Beyond sex-role stereotypes*. Boston: Little, Brown, 1976.
- Bry, A. *The sexually aggressive woman*. New York: Wyden, 1975.
- Hunt, H. *Sexual behavior in the 1970's*. New York: Playboy Press, 1974.
- Liddick, B. Troubled sex-perform or panic. *San Francisco Chronicle*, August 18, 1972, pp. 23; 28.
- Masters, W. H., & Johnson, V. E. *Human sexual response*. Boston: Little, Brown, 1966.
- Spence, J. T., & Helmreich, R. The Attitudes Toward Women Scale. An objective instrument to measure attitudes toward the rights and roles of women in contemporary society. *JSAS Catalog of Selected Documents in Psychology*, 1972, 2, 66-67. (Ms. No. 153)
- Zeldow, P. B. Psychological androgyny and attitudes toward feminism. *Journal of Consulting and Clinical Psychology*, 1976, 44, 150.

Received October 11, 1977 ■

Variations in a Construct: Quantitative and Qualitative Differences in Children's Locus of Control

Philip C. Kendall
University of Minnesota

A. J. Finch, Jr.
Virginia Treatment Center for Children
Richmond

Verda L. Little and Bernard M. Chirico
Virginia Commonwealth University

Thomas H. Ollendick
Indiana State University

The present study examined differences in locus of control scores and factor patterns among normal and nonnormal groups and male and female normals using the Nowicki-Strickland Locus of Control Scale for Children. Multidimensionality of the locus of control construct in children was supported by separate factor analyses of emotionally disturbed, delinquent, and elementary public school children. Qualitative differences in factor patterns between normals and nonnormals raised questions about interpreting inventory scores as reflecting the same construct for differing subject groups.

Locus of control refers to whether people perceive positive and negative outcomes of events as being contingent on their own behavior (internal) or the result of luck, fate, or powerful others (external). Considerable research has been conducted to investigate locus of control as a generalized expectancy (see Lefcourt, 1972), and factor analytic studies with adults support a multidimensional conceptualization of the locus of control construct (Levenson, 1973).

The present studies were designed to examine locus of control as it operates in normal (N1) and emotionally disturbed (ED) children and in juvenile delinquents (JD). Data from a second and larger group of normals (N2) were collected to allow an examination of sex differences.

The N1 group (M age = 10.9) contained 64 male and 43 female children. The N2 sample included 145 males (M age = 10.3) and 125 females (M age = 10). Although obtained from different systems, both normal groups were from suburban public schools in middle-class neighborhoods. The ED group contained 189 hospitalized children (M age = 11.1; 151 males, 38 females) for whom psychiatric screening indicated average

or above intellectual potential and a wide range of diagnoses. Although chronologically older (M age = 15.3), the JD group of 185 children (144 males, 41 females; M IQ of 69 children = 83.5) had a mean reading level (4.5) and mental age (11.3) comparable to children in the other groups.

All children completed the Nowicki-Strickland Locus of Control Scale for Children (Nowicki & Strickland, 1973) with individual attention provided when necessary to insure understanding. This scale was selected over other measures, because it has been found to be unrelated to social desirability or intelligence test scores and because it was thought to be the most reliable measure of generalized locus of control appropriate for children of a variety of ages. Scales were scored in the external direction.

The respective means and standard deviations for the N1, ED, and JD groups were 14.63, 5.59; 17.54, 4.41; and 15.60, 5.17.¹ Results of t -test comparisons between independent means indicated that each of these three groups differed from one another. More external than the N1 group were both the ED sample, $t(294) = 4.65$, $p < .001$, and the JD group, $t(290) = 2.23$, $p < .05$. The ED group also scored in a more external direction than the JD group, $t(372) = 3.93$, $p < .01$.

Portions of the present article were conducted while the first author was affiliated with Virginia Commonwealth University.

Requests for reprints should be sent to Philip C. Kendall, Department of Psychology, Elliott Hall, 75 East River Road, University of Minnesota, Minneapolis, Minnesota 55455.

¹ Though the age-appropriate grade means reported by Nowicki and Strickland (1973) show a marked change from the fifth to sixth grade, the means of the present normal samples, also at these grade levels, were reasonably comparable.

Separate factor analyses resulted in eight factors for the N1 group (accounting for 67% of the total variance), five for the ED (59.2%), and six for the JD group (62.1%). The first factor emerging from the N1 group analysis contained four items that could appropriately be labeled *Generalized Expectancy*. A similar factor did not emerge from the ED or the JD data. Rather, the first ED factor was labeled *helplessness*, and the first JD factor was called *Superstition*. Representative of the other N1 factors were Intellectual Concern and Effort, and other ED factors were Persecution, Superstition, and Futility. The JD factors included Helplessness at Home, Helplessness with Friends, and Helplessness with Parents. Since the three analyses were rotated to the same varimax criterion, emergence of a general factor within the normal children's responses but not in the nonnormal groups should be attributable to differential characteristics in the subject groups.

A separate analysis of the amount of variability accounted for by each factor was conducted to compare the relative potency of factors in each group. In the N1 group, the factors did not account for different amounts of variance (largest $s = 1.66$). In contrast, the first factor to emerge in each of the nonnormal analyses (i.e., Helplessness from the ED data and Superstition from the JD data) accounted for a significantly greater amount of variance ($z = 2.16$, $p < .05$; $z = 2.02$, $p < .05$, respectively) than did the second or remaining factors.

Since whether a child is male or female in our culture is important regarding expectancies, a second sample of normal children (N2) that contained a large number of both males and females was obtained and analyzed for sex differences. A t test between independent means indicated that the females responded significantly more externally, $t(268) = 1.98$, $p < .05$, than males. In addition, both the entire N2 sample and the N2 females were more external than the N1 group, $t(376) = 2.87$, $p < .01$; $t(230) = 4.76$, $p < .01$, respectively, and the JD group, $t(454) = 2.00$, $p < .05$; $t(308) = 3.81$, $p < 0.1$, respectively, but not from the ED group (both $ts < 1$). The N2 males also were significantly more external than the N1 group, $t(250) = 3.12$, $p < .01$, and the JD group, $t(328) = 2.05$, $p < .05$, but, unlike the females, N2 males were significantly less external than the ED group, $t(332) = 1.71$, $p < .05$.

When the total N2 group responses were factor analyzed, the major factor—as in the initial N1 analysis—was appropriately labeled *Gen-*

eralized Expectancy. The remaining factors, however, consisted of only two items and could not be meaningfully labeled. The separate factor analyses for males and females resulted in seven and eight factors, respectively. For males, the major factor consisted of four items concerning Parental Fairness. For females, the major factor consisted of only two items (as did six of the eight factors) and was inconsistent regarding labeling. In regard to relative factor potency, the first male and female factors did not account for significantly different amounts of variance than the remaining factors.

Results of the initial group comparisons support the adjustment—locus of control hypothesis—that internals are better adjusted—since normals were significantly more internal than the two nonnormal groups. The N2 group, however, was significantly more external than the N1 group and not different from the ED group. Although N2 females were more external than the N2 males, one cannot explain the differences between the two normal groups on the basis of sex, since the percentage of females was similar in each group (N1 = 40%, N2 = 46%), and N2 males as a group were more external than the N1 group, composed of both males and females. Though sampling may account for the findings, the most parsimonious explanation may be that locus of control reflects relevant differences in children's life situations. That is, children's scores may indicate some effects of the situation on immediate responses to the scale (i.e., a state) as well as a stable disposition (i.e., a trait). To the extent that this were true or that other unknown factors were operating, one could anticipate that locus of control research with seemingly "normal" groups of children might yield contradictory findings.

In general, although the present results require replication, factor analyses indicated that locus of control in children is a multidimensional concept and that the patterning of factors differs among groups. One clear difference that emerged was in the *themes* that constituted the factors. For example, the main factor in both normal groups was a generalized expectancy theme, whereas the majority of the ED factor labels reflected feelings (e.g., Helplessness, Persecution, Futility), and the JD factors emphasized situations or environments (e.g., at Home, with Peers, with Parents).

Based on these findings, operation of the locus of control construct in different children's groups may be said to differ qualitatively. It seems pertinent then to question whether these qualita-

tive differences could affect quantitative comparisons. Or, stated differently, to what extent is one entitled to interpret scores on a particular inventory as reflecting the same construct for differing subject groups, especially when one is a nonnormal group?

References

Lefcourt, H. M. Recent developments in the study of locus of control. In B. A. Maher (Ed.), *Pro-*

gress in experimental personality research (Vol. 6). New York: Academic Press, 1972.

Levenson, H. Multidimensional locus of control in psychiatric patients. *Journal of Consulting and Clinical Psychology*, 1973, 41, 397-404.

Nowicki, S., & Strickland, B. A. A locus of control scale for children. *Journal of Consulting and Clinical Psychology*, 1973, 40, 148-154.

Received November 11, 1977 ■

Therapist and Group Contact as Variables in the Behavioral Treatment of Obesity

Kelly D. Brownell, Carol L. Heckerman, and Robert J. Westlake
Brown University/Butler Hospital

Obese females were randomly assigned to one of three experimental conditions: (a) a "standard" behavioral treatment (SBT) group emphasizing self-management techniques (Subjects attended group therapy meetings weekly for 10 weeks, then monthly for 6 months and were given a weight control manual.); (b) a group receiving the weight control manual via mail with little professional contact (MMC); and (c) a waiting list control condition. Results revealed a superiority of both treatment conditions over the control condition at posttreatment. SBT subjects did significantly better than MMC subjects at posttreatment but not at the 6-month follow-up. Weight loss for MMC subjects was minimal. The use of "do-it-yourself" treatment manuals is challenged.

The publication of do-it-yourself weight reduction books and manuals has proliferated in recent years. Unfortunately, the distribution of these materials has preceded experimental justification for their use.

Hagen (1974) found a bibliotherapy condition (manual via mail) to be as effective as a behavioral group therapy condition, both of which were superior to no treatment. However, a short follow-up and the use of a mildly obese population make the results difficult to interpret. Hanson, Borden, Hall, and Hall (1976) replicated Hagen's findings but found no treatment effects at a 1-year follow-up. The authors noted that only 4 of 38 subjects attained even 50% of their desired weight loss.

Fernan (Note 1) found that do-it-yourself dieters receiving little professional contact did not differ from subjects receiving no treatment. The present study was designed to remedy the problems with earlier investigations and to evaluate the effectiveness of self-administered diets.

Subjects were 29 females, at least 15 pounds (6.8 kg) or 15% overweight, who were randomly assigned to a standard behavioral treatment group (SBT), a group receiving a manual with minimal professional contact (MMC), or a no-treatment group. Subjects averaged 63.6% overweight, mean weight was 194.4 pounds (88.4 kg),

and average age was 48.7 years. In the SBT group subjects met weekly with a trained therapist for 10 weeks, then monthly for 6 months and were given the *Behavioral Weight Control Manual* (Brownell, Heckerman, & Westlake, Note 2). MMC subjects received the same manual and met six times to be weighed.

There were no pretreatment differences among groups for mean body weight or mean percentage overweight. At posttreatment, differences among groups were significant for change in body weight, $F(2, 25) = 10.36, p < .01$; change in percentage overweight, $F(2, 25) = 11.62, p < .01$; and the weight reduction quotient (pounds lost/pounds over $\times 100$), $F(2, 25) = 12.49, p < .01$. Newman-Keuls comparisons revealed that SBT and MMC subjects were superior to control subjects and that SBT subjects lost more weight than MMC subjects for all three measures of weight change ($p < .05$). At the 6-month follow-up, there were no differences between groups. Mean weight loss was 7.42 pounds (3.7 kg) for SBT subjects and 2.2 pounds (1.0 kg) for MMC subjects.

Both treatments were plagued by a lack of long-term effectiveness, and the weight losses could be considered only temporary. This was particularly true of the minimal contact group, which failed to produce even temporary weight loss. These results are consistent with previous findings that self-administered diets do not produce sustained weight loss.

The emotional hazards of unsuccessful dieting have been clearly documented (Stunkard & Rush,

Requests for reprints and for an extended version of this study should be sent to Kelly D. Brownell, who is now at the Department of Psychiatry, University of Pennsylvania, 205 Piersol Building, Philadelphia, Pennsylvania 19104.

1974), and there is preliminary evidence which suggests that periods of weight gain are characterized by substantial increases in blood pressure and serum lipid levels (Gordon & Kannel, 1973). It is possible that an ineffective diet is more dangerous than no diet.

In light of these potential hazards, and considering that the do-it-yourself dieter has little medical or psychological guidance, diets designed to be self-administered should be subject to controlled clinical investigation prior to distribution, and consumers should be educated as to the merits and drawbacks of specific programs.

Reference Notes

1. Fernan, W. S. *The role of experimenter contact in behavioral bibliotherapy of obesity*. Unpublished manuscript, Pennsylvania State University, 1973.
2. Brownell, K. D., Heckerman, C. L., & Westlake,

R. J. *Behavioral Weight Control manual*. Unpublished manuscript, Brown University, Butler Hospital, 1975.

References

- Gordon, T., & Kannel, W. B. The effects of overweight on cardiovascular disease. *Geriatrics*, 1973, 28, 80-88.
- Hagen, R. L. Group therapy versus bibliotherapy in weight reduction. *Behavior Therapy*, 1974, 5, 222-234.
- Hanson, R. W., Borden, B. L., Hall, S. M., & Hall, R. G. Use of programmed instruction in teaching self-management skills to overweight adults. *Behavior Therapy*, 1976, 7, 366-373.
- Stunkard, A. J., & Rush, J. Dieting and depression reexamined: A critical review of reports of untoward responses during weight reduction for obesity. *Annals of Internal Medicine*, 1974, 81, 526-533.

Received December 5, 1977 ■

CHARLES C THOMAS • PUBLISHER

COUNSELING IN COMMUNICATIVE DISORDERS edited by R. E. Harbauer, *Andrews Univ., Berrien Springs, Michigan*. (13 Contributors) This book comprehensively covers the psychological and emotional problems of persons with communicative disorders. Each chapter focuses on a particular communicative problem, giving a general overview of the disorder, counseling techniques and procedures, and suggestions for evaluation of effectiveness. Organic and functional disorders, the hearing impaired, financial aid, and other topics are explored. '78, 336 pp., 2 il., 1 table, \$16.50

ASSERTIVE TRAINING FOR WOMEN (2nd Ptg.) by Susan M. Osborn, *Bellevue Community College, Bellevue, Washington*, and Gloria G. Harris, *Univ. of Washington, Seattle*. The authors review ways in which women are socialized to be submissive and dependent and they evaluate the usefulness of traditional and alternative psychotherapeutic approaches. Specific areas of application explored include improvement of social interactions, enhancement of sexual functioning, entrance and advancement in the job market, and relief of emotional depression. '78, 216 pp., 12 il., 7 tables, cloth-\$10.50, paper-\$7.95

THE CLINICAL APPLICATION OF PROJECTIVE DRAWINGS (5th Ptg.) edited by Emanuel F. Hammer, *New York Univ., New York City*. (14 Contributors) This book provides a grounding in the fundamentals of the subject and goes on to consider differential diagnosis, psychodynamics, conflict and defense, psychological resources as treatment potentials, and projective drawing usage in therapy. '78, 688 pp., 360 il., \$21.75

PUPIL PERSONNEL STRATEGIES AND SYSTEMS by Robert W. Ferris, *Cherry Hill Public Schools, Cherry Hill, New Jersey*. This comprehensive text bridges the gap between the theoretical and the practical in the pupil personnel field. Chapters are included on management models, administrative systems, group counseling, career development, use of computers, and assessment and evaluation. Psychologists, counselors, administrators, social workers, learning specialists, and related professionals will all benefit from the depth and clarity of this thorough examination of pupil personnel strategies and systems. '75, 264 pp., 22 il., 5 tables, \$14.95

STRATEGIES FOR HELPING STUDENTS by Calvin D. Catterall, *Staff Development Consultants, Columbus, Ohio*, and George M. Gazda, *Univ. of Georgia, Athens*. All available strategies for helping students are presented. They are grouped into four areas of concentration: strategies for developing academic/intellectual skills, corrective/remedial interventions, strategies for developing humanistic/life skills, and corrective/therapeutic interventions. The final chapter explains the process of selecting the strategy most likely to succeed with a specific need. '78, 416 pp., 5 il., 38 tables, \$18.75

PERSONALITY PROJECTION IN THE DRAWING OF THE HUMAN FIGURE: A Method of Personality Investigation (10th Ptg.) by Karen Machover, *Dept. of Hospitals, New York City*. This original technique for human figure drawing analysis combines the psychology of expressive movement with a fresh approach to functional and symbolic significance. Illustrative case studies demonstrate the manner of integrating the data to form a coherent picture of the personality. '78, 192 pp. (5 1/2 x 8 1/2), 18 il., \$8.50

THE PSYCHOLOGICAL REHABILITATION OF THE AMPUTEE by Lawrence W. Friedmann, *Nassau County Medical Center, East Meadow, New York*. This text begins with an historical and cultural analysis of amputation and its effects on the individual. Other chapters focus on patient reactions, prostheses, social aspects, the phantom limb phenomenon, the limb-deficient child, and the occurrence and treatment of phantom pain. Analyses of psychological responses to amputation are presented in this book with unprecedented depth and insight. '78, 176 pp., 13 il., \$13.50

DIAGNOSIS AND REHABILITATION IN CLINICAL NEUROPSYCHOLOGY by Charles J. Golden, *The Univ. of South Dakota, Vermillion*. This book offers a unique integration of the theoretical basis, experimental research, and clinical and diagnostic procedures of the emerging field of clinical neuropsychology. The main portion of the text is concerned with clinical test procedures and with ways in which psychological testing can be used to identify a brain disorder, localize the area of a lesion, and determine underlying causes. '78, 304 pp., 11 il., 13 tables, \$14.75

Orders with remittance sent, on approval, postpaid

Evaluation of Psychological Service Delivery Programs



Special Issue of Professional Psychology

Evaluation of Psychological Service Delivery Programs is the title of a special November, 1977 issue of Professional Psychology. Editor Allan Barclay and guest Co-editors Bob and Evelyn Perloff have assembled a distinguished collection of articles that brings readers up to date on the evaluation of psychological services. The issue addresses the need to justify the effectiveness of service delivery programs not only in terms of cost but more importantly in terms of human benefits.

Sample Topics and Contributors

A sample of the topics and contributors appearing in this issue includes: Theodore Blau on quality of life; Nicholas Cummings on psychotherapy; Howard Blane on alcoholism; S. B. Sells and associates on drug abuse, and NIMH's Bertram Brown, Benjamin Liptzin and James Stockdill on the federal view of mental health program evaluation.

Ordering Information

Single copies of the issue are available from the Order Department at APA for \$5 each. All orders \$25 or under must be prepaid. Please make checks payable to APA. Send your orders to:



Order Department, P.P.
American Psychological
Association
1200 17th Street, N.W.
Washington, D.C. 20036

Chicago Conference on the Psychology of the Self

October 13, 14, & 15
1978
Ritz-Carlton Hotel
Chicago

A recent scientific event in the fields of psychology and psychiatry has been the emergence of the Psychoanalytic Psychology of the Self. In order to acknowledge and appraise the significance in psychoanalysis of this new development, a conference will bring together a group of mental health professionals from the United States and Europe. The conference will examine self psychology, identify its current problem areas, and outline plans for further development and applications. The conference program consists of six symposia with discussants and related workshops, dealing with the following topics:

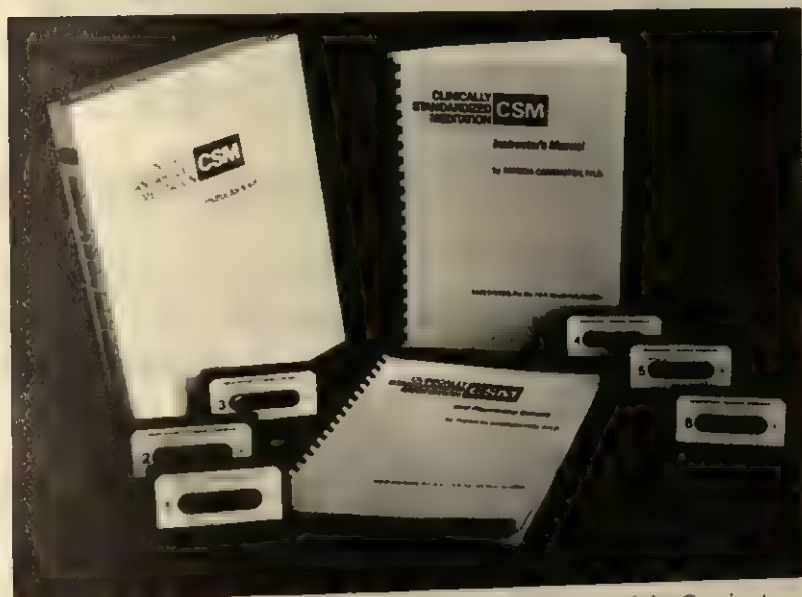
- ☆ Self Psychology and the Concept of Health
- ☆ Does the Psychology of the Self Narrow the Gap Between Psychoanalysis and the Other Sciences of Man?
- ☆ Self Psychology and the Concept of Psychopathology: A Case Presentation
- ☆ Psychoanalytic Developmental Psychology: An Integration of Theories
- ☆ Psychoanalysis of the Self and Psychotherapy
- ☆ Concluding Remarks and Future Directions

The Conference Planning Board is composed of members of the Institute for Psychosomatic and Psychiatric Research and Training of Michael Reese Hospital and Medical Center, and the Institute for Psychoanalysis in Chicago. Please note that attendance will be by advance registration only, and that registration will be limited.

For registration information, write:

Chicago Conference on the Psychology of the Self
Psychosomatic and Psychiatric Institute (P&PI)
Michael Reese Hospital and Medical Center
29th Street and Ellis Avenue
Chicago, Illinois 60616

ANNOUNCING A BREAKTHROUGH IN THE MEDITATIVE-RELAXATION FIELD...



*CSM Instructor's Kit: Created and narrated by Patricia Carrington, Ph.D., Dept. of Psychology, Princeton University (the technique described in her book **FREEDOM IN MEDITATION**)*

***CONTENTS:** 3 Meditation training tapes; 3 clinical lectures; 82 page semi-programmed-instruction text; instructor's manual.*

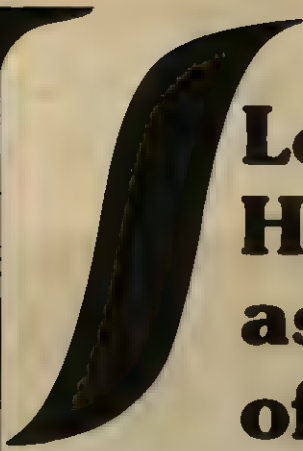
A fully automated, self-regulating course covering all elements involved in Clinically Standardized Meditation (CSM), motivational tapes, exercises, reinforcement aids and supervisors' guidance.

"Superbly done. . . highly effective. . . will be a major tool for clinicians and researchers."

Arnold A. Lazarus, Ph.D.

For free brochure write to:

PACE EDUCATIONAL SYSTEMS INC.
P.O. Box 113-A
Kendall Park, New Jersey 08824



Learned Helplessness as a Model of Depression

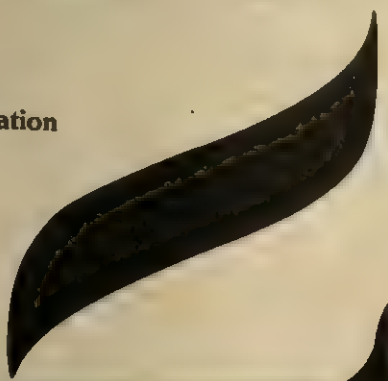
**Special issue of the
Journal of Abnormal Psychology,
February 1978.**

The issue serves as a forum to investigate the validity and adequacy of the learned helplessness model of depression in humans and includes both critical and supportive theoretical and empirical articles with comments.

L. Rowell Huesmann is guest editor. Richard A. Depue, Martin E. P. Seligman, Charles G. Costello, Robert C. Smolen, and Alexander M. Buchwald are among the contributors.

Copies of the special issue are available at \$6 each, prepaid. Discounts on bulk orders are also available. To order single copies or for more information on bulk orders write:

**Subscription Department
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036**



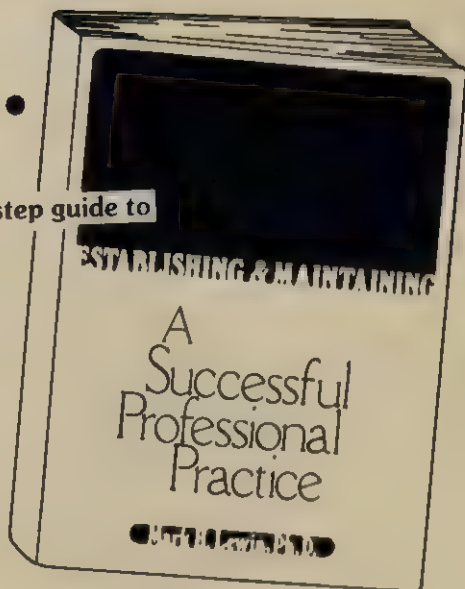
For the First Time...

an easy reading, practical, step by step guide to

A must for

- Full and part-time private practitioners.
- Those who have ever even considered entering private practice.

"Reduces the chance of failure or disappointment."



An entertaining, well organized book complete with worksheets, plans, and practice approaches to the everyday "nuts and bolts" business aspects of private practice.

MARK H. LEWIN, an ABPP Diplomate, distinguished author and editor, and founder of a thriving independent psychological service center, gives step-by-step procedures for:

- Assessing your readiness for private practice
- Preplanning the practice
- Setting up the office
- Developing new referrals
- Managing your time
- Effective use of lawyers and accountants
- Evaluating the practice as it develops
- Budgeting, fee setting, controlling costs and tax expenses

First Printing 1978 - \$25.00
10% Discount Available Now
\$22.50

Mail checks payable to:
Professional Development Institute
756 East Main Street
Rochester, N.Y. 14605

Please send me _____ copy(ies) of **Establishing and Maintaining a Successful Professional Practice** for which I enclose \$22.50 per book plus New York sales tax where applicable.

Name _____

Street Address _____

City _____ State _____ Zip _____

JCC

PSYCHOLOGICAL TEST SPECIALISTS

Exclusive publishers and distributors of selected psychological tests

Quick Test (QT)

A very quick and satisfactorily reliable screening test of intellectual ability (8 to 12 minutes total time in all three forms). Words are defined non-verbally by choice among drawings. Suitable for any person (no matter how physically disabled) who can signal a yes-no response in any way. Covers intelligence levels from two years to superior adult. Excellent for building rapport, and for quick testing before and after any procedure. \$16.00 for set of materials

Memory-For-Designs Test (MFD)

Simple drawing test of perceptual-motor coordination, serving as extremely sensitive detector of brain injury of many types. \$17.00 for set of materials

Group Personality Projective Test (GPPT)

Multiple-choice questions about ambiguous stick-figures make possible rapid and effective measurement of significant, factor-analytically isolated aspects of personality. \$27.00 for set of materials

Id-Ego-Superego Test (IES)

Subtests give independent estimates of personality components from a modified psychoanalytic point of view. \$43.50 for set of materials

Kahn Test of Symbol Arrangement (KTSA)

Uses symbolic plastic objects which are arranged and sorted; validly discriminates among normals, neurotics, brain-injured, and psychotics, and indicates accompanying personality dynamics. \$50.00 for set of materials

Full-Range Picture Vocabulary Test (FRPV)

Brief test of intelligence in two forms using the same 16 plates. Words are defined by choosing pictures. Valuable for intelligence screening, and with the physically handicapped. \$15.00 for set of materials

Proverbs Test (PT)

Group and individual forms determine similarity of thought processes to those of various normal and clinical groups. Highly sensitive indicator of psychotic processes. \$13.00 for kit

Vocational Apperception Test: Advanced Form (VAT:ADV)

Separate series of plates showing male and female college-rated occupational activities, accompanied by explanation of use and interpretation as a projective test. \$14.50 for set of materials

Famous Sayings (FS)

Agreement with proverbs, aphorisms, and folk sayings is analyzed to determine personality structure. Particularly useful in personnel selection. \$13.00 for kit

It Scale For Children (ITSC)

An experimental picture scale for measuring sex role preference in children. \$30.00 for set of materials

Kahn Intelligence Tests: (KIT:EXP)

Relatively culture-free intelligence tests requiring no reading or writing. Supplementary scales for assessment of intelligence of the blind, ability in concept formation, recall, and motor coordination. One scale can be administered by sign language. \$52.00 for set of materials

PSYCHOLOGICAL TEST SPECIALISTS

Box 9229

MISSOULA, MONTANA 59807

VOLUME 16, NUMBER 4

AUGUST 1978

SPECIAL ISSUE: METHODOLOGY
IN CLINICAL RESEARCH

JOURNAL OF
CONSULTING
AND
CLINICAL
PSYCHOLOGY

AMERICAN PSYCHOLOGICAL ASSOCIATION
ISSN 0022-006X

INSTRUCTIONS TO AUTHORS

Style of manuscripts. Manuscripts must be prepared in the style described in the *Publication Manual of the American Psychological Association* (2nd ed.). Instruction on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors should submit manuscripts in triplicate and should keep a copy to guard against loss.

Abstracts. Manuscripts of regular articles must be accompanied by an abstract of 100-175 words. Manuscripts of brief reports must be accompanied by an abstract of 75-100 words. All abstracts must be typed on a separate sheet of paper.

Brief Reports. The *Journal of Consulting and Clinical Psychology* will accept Brief Reports of search studies in clinical psychology. The procedure is intended to permit the publication of soundly designed studies of specialized interest or limited importance which cannot now be accepted as regular articles because of lack of space. Several pages in each issue will be devoted to Brief Reports, published in the order of their receipt without respect to the dates of receipt of the regular articles. Most Brief Reports appear in the first or second issue to go to press following their final acceptance.

An author who wishes to submit a Brief Report:

1. Sends the Brief Report, limited to two printed pages and prepared according to the specifications given below.

2. Agrees not to submit the full report to another journal of general circulation.

Specifications. The Brief Report should give a clear, condensed summary of the procedure of the study and as full an account of the results as space permits.

To ensure that a brief report does not exceed two printed pages, follow these instructions for typing: (a) Set typewriter to a 48-space line. (b) Type text. (c) Count all lines except abstracts, title, and by-line, including acknowledgments. If you have exceeded 185 lines, shorten the material. (d) If your brief report barely exceeds 75 lines (one printed page), try to edit to 75 lines exactly. Reports slightly exceeding one page will be edited for length in order to use space economically.

In Brief Reports, headings, tables, and references are avoided or, if essential, must be counted in the 75 lines. This journal no longer requires an extended report, but, if one is available, each Brief Report must be accompanied by the following footnote:

Requests for reprints and for an extended report of this study should be sent to John Doe (give the author's full name and address).

The footnote should be typed on a separate sheet and *not* counted in the 75-line quota.

Garfield Appointed Editor, 1979-1984

The Publications and Communications Board of the American Psychological Association announces the appointment of Sol L. Garfield as Editor of the *Journal of Consulting and Clinical Psychology* for the years 1979-1984. As of January 1, 1978, manuscripts should be directed to the Editor-elect:

Sol L. Garfield
Department of Psychology
Washington University
St. Louis, Missouri 63130

Preface

Four of every five manuscripts submitted to this journal in recent years have been rejected. Although rejection can occur for many reasons, most of the unsuccessful manuscripts were returned because of serious methodological weakness in the research. Many of these errors of method were fundamental and glaring. Indeed, it is not possible to complete an editorial term without inferring reluctantly that current doctoral training in research, as reflected by recent submissions to this journal, is deficient to the point of being disastrous.

This unfortunate inference has led to this special issue, which is intended to be a resource for behavioral scientists working on problems of clinical psychology. Each contribution that follows describes and discusses the major methodological aspects of a particular topic area such as smoking or addiction. The topic areas have been selected because of the frequency with which they have stimulated contemporary research investigation.

Readers may note that at some points contributors have raised issues that may appear elementary. This has been done not out of an insensitivity to the level of sophistication of the journal's readers, but because our experience reviewing manuscripts has made it clear that many investigators are misinformed on some basic matters.

It is the hope of the Editor and contributors alike that this issue of the journal will be of some lasting value to those who engage in the difficult and demanding task of clinical research.

Brendan A. Maher
Editor

Diary No.	31
Date	30.7.4.18
at	Leeds
Bureau Econ. & Soc. Research.	

Research Problems in Clinical Diagnosis

Sol L. Garfield

Washington University in St. Louis

This article discusses some of the problems and deficiencies apparent in past research on clinical diagnosis. Among the important issues discussed are those pertaining to sampling problems with regard to the clinical subjects studied, adequacy of control groups, base rates, clinical versus statistical significance, lack of cross-validation, and problems due to reliance on inadequate schemes of classification. Some cautions and suggestions in terms of future research are also discussed.

Although interest among clinical psychologists in clinical diagnosis and related activities has appeared to diminish relatively in recent years, interest in diagnostic problems remains, and a number of studies on such problems are conducted yearly. Although over the years we have had an opportunity to become better acquainted with methodological problems concerning research in this area, it would appear as if rather limited advantage has been taken of this opportunity. I will discuss a number of methodological inadequacies that I have encountered in reviewing manuscripts for this and other journals, as well as in the published literature.

The term *clinical diagnosis* does not appear to have a precise meaning, and consequently, this should be acknowledged at the outset. Although in the past, this designation frequently has referred to psychiatric diagnosis, or a nosological label taken from the *Diagnostic and Statistical Manual* prepared by the American Psychiatric Association (1968), such a delimitation has not been exclusive. Many psychologists have not limited themselves in this manner, preferring to use other schemes, formal or informal. Some have preferred the term *assessment* to that of *diagnosis*. Others have preferred to rely on personality descriptions and on the inferred psychodynamics of the individuals appraised.

I will not debate the issue here of what clinical diagnosis really is, but I will include in my discussion references to research that involves the clinical assessment, diagnosis, comparison, or categorization of individuals being appraised or studied for clinical purposes. The presentation is organized around several topics that have appeared to be of particular concern.

Sampling Problems

One of the most frequent and serious problems encountered in diagnostic research pertains to the sample of subjects or patients used in particular investigations. This actually is a broad categorization for a variety of problems. One critical issue pertains to how far one can generalize from the findings obtained with a certain sample of subjects purported to represent a given type of disorder or diagnosis. Ostensibly, at least, findings secured from a particular sample representative of a given category of disorder presumably have relevance for other comparable samples. The question then is, how does one define both the sample studied and the other samples or populations to which the results are supposedly applicable?

Unfortunately, there are no true standard reference measures that can be used to define a particular clinical group of subjects, and the diagnostic terms used leave much to be desired. Consequently, it is exceedingly important that the sample be selected with great care and that as much useful descrip-

Requests for reprints should be sent to Sol L. Garfield, Department of Psychology, Washington University, St. Louis, Missouri 63130.

tive information as possible be provided concerning the sample. Many manuscripts, as well as a significant number of published articles, fail to provide even such basic data as the sexual composition of the sample, let alone other attributes of importance. Such aspects as age, length of hospitalization, frequency of hospitalization, marital status, work history, previous treatments, family resources, education, intelligence, type of ward the patient is on, whether he or she is receiving medication, the type of medication, and the like, are all potentially important variables that may influence test performance, treatment outcome, and similar variables. It should be apparent that significant variation on some or many of these variables between studies limits the reliability of the results secured and the drawing of conclusions that may have broad applicability.

The problems referred to above are particularly visible when samples of modest size are drawn from institutional settings that vary widely on a number of dimensions. Patients in university or private hospitals generally are quite different from those in state or Veterans Administration psychiatric hospitals, and generalizations from one setting do not fit the other settings, even though the patients may all carry a diagnosis of schizophrenia. For example, in a previous study of prognostic scales used in research on schizophrenic patients, we were unable to secure an adequate number of "reactive" or good-prognosis patients in the state hospital where we were conducting our study and had to secure patients from a city hospital, which had fewer chronic patients (Garfield & Sundland, 1966).

Although sample size is also of some importance, since small scale studies are more prone to produce findings limited in reliability, size alone is not sufficient. The selection of subjects and sample specifications is clearly of prime importance. If one is studying or comparing selected types of disordered behavior, the criteria used in selecting subjects should be explicit and the procedures used should be ones for which the reliability and validity are known or available, and meet commonly accepted standards. Particularly when studies are conducted on groups based on psychiatric diagnosis, it is essential that

how the diagnoses were derived be clearly described and that other supporting selection criteria be used. Diagnoses based on old records and provided by different psychiatrists are usually not sound or reliable bases for subject selection.

A related problem concerns the matter of randomization of subject selection. Were the subjects selected at random from a previously selected pool of available subjects or were they selected because they were not on drugs, were in a special ward, were considered co-operative, or in some other manner were really not "typical subjects"? Such selectivity can obviously bias the results obtained and can limit their generalizability.

Selection of subjects on the basis of a single scale may also not be adequate, particularly if the subjects are considered to represent a particular diagnostic group. For example, not all subjects who score 70 or higher on a scale of the Minnesota Multiphasic Personality Inventory (MMPI) resemble groups diagnosed on other criteria as schizophrenic, depressed, and so on. College students who secure such scores may or may not be clinically depressed, psychotic, and so forth, and comparisons with actual clinical populations may therefore not be warranted.

It would seem clearly desirable to use more than one procedure or method to establish the diagnosis of the subjects to be used in any research study in which the diagnosis is considered to be a significant variable. Psychiatric or clinical diagnosis should be supplemented by other criteria such as scores on appropriate tests or standardized rating scales. In the case of depression, for example, scores on the MMPI, the Hamilton Rating Scale (Hamilton, 1960), and the Beck Depression Inventory (Beck, 1972) could be used in addition to clinical diagnoses. Furthermore, clinical diagnoses, preferably, should be based on diagnoses secured from two or more clinicians with a reasonably high indication of reliability between them. If such procedures are followed, there is greater assurance of the reliability of diagnosis as well as several external or operational reference points in defining the samples used. To be sure, extra effort may be required to carry out such procedures, and, to a certain extent,

such procedures might reduce the size of potential samples and raise new issues of selectivity and limited generalizability. However, the samples would be more clearly defined, and the dangers of relying exclusively on somewhat haphazard means of classification would be lessened. In the long run, at least one would hope, the conflicting results obtained with the use of diffuse, unclear, and unreliable diagnostic categorization might diminish.

Another issue that is sometimes apparent concerns the selectivity of a patient sample in order to match it with a control group. If, for example, the control group has an average IQ of 98, an average educational attainment of 10th grade, or is composed primarily of ward attendants, a subject sample selected to match it on one or more of these variables may be a highly selected and unrepresentative sample of the patient group that these people are supposed to represent. They are not typical patients of a particular diagnostic group, and broad generalizations to other more randomly selected groups are not justified. For example, in one study of mine of patients diagnosed as schizophrenic, very different patterns of performance on the Wechsler-Bellevue Scale were secured for patients differing in education and IQ (Garfield, 1949). Furthermore, comparisons of samples of schizophrenic patients studied in different investigations, which differed noticeably in terms of mean IQ, also revealed significant differences on test patterns among these samples. In other words, there were as many significant differences between the various samples of schizophrenic patients as there were between a given sample and a normal group of control subjects. A recent study of brain-damaged adults also revealed significant differences due to education on a battery of tests (Finlayson, Johnson, & Reitan, 1977).

The relative influence of social class variables on samples of subjects is another matter that needs to be carefully appraised in each study. As Meehl (1971) has pointed out, the tendency among social scientists to view correlations uncorrected for social class as automatically spurious is unjustified, unless a clear case can be demonstrated that a causal relationship exists between social class and the

correlated variables in the specific instance. In some cases, this may be so, but in others, the so-called nuisance variable may be of little importance. If there is doubt, the investigator can provide both corrected and uncorrected correlations as one means of attempting to evaluate the possible importance of social class variables. (See Meehl's, 1971, article for a more extended treatment of this issue.)

The selection and specification of subjects used in research on clinical diagnosis is thus a matter of primary importance, and would-be investigators should give careful attention to the issues discussed in the preceding paragraphs. In the final analysis, the results can be no better than the type and representativeness of the subjects used.

Some of the problems pertaining to how clinical diagnoses are secured will also be mentioned briefly here, although they could be accorded a separate section. Most clinical psychologists are quite aware of the problem of diagnostic reliability, and this issue need not be reviewed in any depth here (Garfield, 1974). However, because of its relevance to much research on clinical diagnosis, some reference to this matter is pertinent here.

In an earlier section of this article, some mention was made of the importance of fully describing the sample studied and how it was selected. One aspect of sample description concerns the matter of diagnosis. How were the patients diagnosed, by whom, and when? These are all pertinent questions that influence sample specification and, ultimately, the results secured and the conclusions drawn. If the diagnosis is made by one clinician, how representative are his or her diagnoses? In one study comparing three psychiatrists working in the same hospital with comparable groups of patients, it was found that one of the psychiatrists classified two thirds of his patients as schizophrenic, whereas the other two psychiatrists classified 22% and 29% of their patients as schizophrenics (Pasamanick, Dinitz, & Lefton, 1959). Thus, some attention has to be paid to the reliability of diagnosis and comparability of diagnoses among different studies.

Diagnoses made at the time of intake or admission to the hospital may also differ from

those made at a later time when more opportunity for observation and study is available. Similarly, diagnoses reached at a formal staff conference or with the help of teaching consultants may be different from those made by a single clinician. Diagnoses based on psychological tests may also be different from those based on other data sources. However, little or no information on such relevant matters is provided in a large number of studies. The usual statement is something like this: "Sixty-four patients diagnosed as schizophrenic were selected for study." This certainly does not provide adequate information about how the diagnoses were made or by whom they were made. Consequently, specification of how diagnoses were reached is of some importance in estimating the confidence one can place in the diagnoses secured and also in evaluating the comparability of research studies. Since the behavior of patients also changes over time, there is also a question about the suitability of diagnoses that were reached some time prior to the current investigation. There are, thus, a number of apparently minor but nevertheless important considerations that pertain to the matter of diagnosis.

Proper Control Groups

Another problem frequently encountered in research reports concerned with clinical diagnosis pertains to the appropriateness of the control groups used. Although we are relatively more sophisticated about such matters now than we were in the past, problems of appropriate controls are still evident. Obviously, a group of patients that has been hospitalized for some time should not usually be compared with a group of apparently normal college students, but comparisons approaching this type have been made.

This problem is of special significance when the focus of a particular investigation is on clinical diagnosis or on the specification of patterns of performance of a specific category of patients or subjects. If the results of the study are to have any clinical significance in the practical sense, then the experimental group must be compared to other clinical groups with which they would normally be

compared in the actual clinical situation. Comparisons of test performance or other measures of a sample of patients diagnosed as schizophrenic should be made with other clinical groups normally seen in that clinical situation and in approximately the usual proportions. In a clinical or hospital setting in which a clinical diagnosis or evaluation is sought, the problem is rarely one of comparing the given patient to a normal population. Rather, the issue is one of differential diagnosis and appraisal. Is there any suggestion of psychotic disturbance or of possible brain damage? How serious is the thought disturbance or depression manifested by the patient? In trying to reach answers to such questions, the clinical psychologist is considering patterns of various types of psychopathology and is not comparing the patient's performance primarily to the performance of normally adjusted individuals. Thus, in studying a diagnostic pattern for possible utility in clinical diagnosis, the investigator should compare the results of the particular clinical group of interest with those of other diagnostic groups that are usually encountered in practice and from whom the aforementioned group is to be compared or differentiated. For purposes of clinical diagnosis, it would seem efficacious to have a control group made up of the proper mix of the other diagnostic groups that are normally encountered in the particular setting or of several groups representing the types of disorders that are most frequently confused with the group under study.

Similar to the issue mentioned earlier with regard to the selection of the sample of subjects for investigation, the investigator should provide adequate information on all groups of subjects, such as how they were selected and from which subpopulation they were drawn. For example, if a control group of 30 patients was selected for purposes of comparison with the group under study, why and how were these subjects selected and from what number of comparable subjects were they drawn? If they were the only ones who had certain data or test scores available, what were the reasons for this particular state of affairs? How selective is the group, and how representative are they of supposedly similar subjects? Selective bias may greatly impair

the kinds of conclusions that may be drawn and the extent to which generalization to other samples and populations is possible.

It should be reasonably obvious also that whatever control group or groups are used should be comparable on the variables of importance for a particular investigation. When cognitive tests are used, the groups should show some comparability in terms of level of ability, education, and age, since these may all affect performance. However, if one's particular focus is the diagnosis of mental retardation, such comparability may not be necessary or meaningful. Length of institutionalization, drugs, type of ward, degree of cooperation, and other such attributes may also be variables of importance.

In essence, therefore, considerable attention should be given to the selection of adequate and appropriate control groups in research on clinical diagnosis. Although this seems particularly relevant to matters of differential diagnosis and related problems, the issue of appropriate control groups also applies to more experimental or theoretical studies of psychopathology—at least as far as I am concerned. Numerous studies of various psychological functions in schizophrenic subjects have compared the latter mainly with normal controls. Although this may have some value at a certain stage of investigation, it is ultimately of limited value if one is interested in demonstrating particular patterns of response or thought in a given clinical disorder. What the investigator is actually trying to discover or demonstrate are response patterns that characterize a pathological group. Whether these patterns are distinctive of the particular type of pathology in question can only be demonstrated if the performance of the pathological group is compared with other pathological groups of comparable severity. In this particular instance, other psychotic groups would appear to be the most adequate control groups.

Clinical Versus Statistical Significance

Surprisingly, a number of basic statistical and clinical considerations appear to get slight consideration from would-be authors and investigators. Those that have been noted with some frequency will be discussed.

In at least a noticeable number of manuscripts, significant correlations or differences at the .05 level of significance are reported, with little attention paid to the number of significance tests performed. Although it should be apparent that the interpretation of the significant findings secured is directly related to the number of statistical tests performed, this stricture is not always observed. If 35 comparisons or correlations are performed and 2 are found to be at the magic .05 level of significance, it seems clear that these results are so close to a chance occurrence that little should be made of them. However, some investigators generally play down this aspect of their results and emphasize their potential significance. In a similar fashion, post hoc analyses are sometimes treated as if explicit hypotheses were being tested. Obviously, different considerations enter into such analyses than when a specific hypothesis, stated in advance of the investigation, is being tested. Although these matters are considered to be rather basic ones in instruction on experimental design and statistical analysis, somehow the lessons on these topics are either not learned well or are quickly forgotten.¹ When investigators emphasize significant findings that they have not predicted in advance but that are noted after the investigation has been completed, they are essentially capitalizing on chance occurrences. For the results to be taken seriously, they should be replicated on a new sample of subjects.

A related issue concerns the practical or clinical significance of findings that are clearly significant statistically at the .05 or even at the .001 level of significance. It would appear from numerous reports of all kinds of studies that researchers have been led to worship at the shrine of statistical significance. Perhaps this is a result of the emphasis placed on securing "positive" results and the dread of not being able to reject the null hypothesis, particularly with reference to doctoral dissertations. Happy is the person who secures "positive results" at the .05 level

¹ These matters along with some procedures for estimating the probability of securing spurious findings are discussed by Hays (1963, pp. 488-489).

of significance, regardless of whether the results are of some potential practical or social value. Of all the limitations that I have encountered in reviewing manuscripts for several journals over a period of years, the emphasis on statistical significance and the disregard of practical psychological significance is probably the most frequently encountered.

There are several aspects of this problem that should be emphasized here even though most of them may be considered to be obvious or elementary. The fact that they occur with some frequency makes such an emphasis permissible and warranted.

Clearly, a researcher must examine his or her data to see if the results obtained are due to chance. A statistical test of the differences secured between clinical groups or of correlations obtained on certain measures and designated criteria for a specific diagnostic category or categories is a necessary procedure to estimate the influence of chance on the results secured. No criticism of this procedure is implied here. However, one should recognize that this represents only one of the necessary procedures for appraising the results. Such a statistical test informs us of the probability that our results may be explained by chance occurrences. If our results are significant at .05 level of significance, our usual interpretation should be that there are only 5 chances in 100 that the obtained results may be attributed to chance. The findings, of course, could be due to chance, but the odds would appear to be against it. However, all one can reasonably conclude is that the findings do not appear to be due to chance and that if the study were repeated, we might expect the findings to be comparable to those obtained initially. Whether the results obtained have any clinical usefulness cannot be determined by the statistical tests alone. Other appraisals must be made for this purpose.

Before going on to discuss the matter of the clinical significance of research data, it is worth reiterating some other elementary considerations pertaining to statistical manipulations and their meaning. Statistical tests tend to be very much influenced by the size of the samples used and by their variability. Obviously, large samples generally would be less influenced by selective and chance vari-

ables than would very small samples, and, thus, findings that are small in actual magnitude may be statistically significant in the former instance, whereas they would fail to attain significance with smaller samples. For example, with samples of about 30 subjects, correlations have to be in the neighborhood of .35 or so in order to reach significance. In contrast to this, with a sample of several hundred subjects, a much lower correlation, in the neighborhood of .10, may be statistically significant, although the amount of predicted variance in the scores is of no practical significance. Thus, besides the level of significance, we must also consider the size of the sample and the actual amount of variance accounted for by the correlation coefficient if we are to interpret the findings in terms of their broader significance or utility. Moderately high correlations that are not statistically significant would appear to be of little value, and highly reliable but low correlations would also appear to be of rather limited value. In a similar fashion, low variability within groups of subjects generally increases the probability of securing statistically significant results, and although small standard deviations offer the possibility of useful discrimination between different clinical groups, the actual utility of the measures or comparisons used requires further analysis of the actual data secured.

Even though there are occasional discussions of the difference between clinical and statistical significance in the published literature (Lick, 1973), the importance of this topic either does not get enough attention during the graduate training of future clinical researchers or the occasional references to it have a relatively weak impact on those to whom it is directed. Many investigators appear content to rest on their statistical laurels and not to worry overly about the actual practical value of their results.

Clearly, as indicated earlier, an investigation can be conducted to test a theoretical proposition or hypothesis when practical considerations are not of primary importance. However, if the investigator in his or her discussion of the results obtained implies some potential clinical significance for the data, then he or she should provide some additional

support for the inferences that go beyond the fact that statistically significant results were obtained. The kind of analyses that should be done will be described briefly in the next few paragraphs.

There are several aspects of the research data that should be examined in terms of their potential clinical significance. For example, regardless of the mathematical procedures used to test hypotheses within a statistical model, it is also important to know how much of the variance is accounted for in terms of the particular variables studied. For practical purposes, as well as theoretical ones, this is a very important consideration, but it is very frequently omitted in the discussion of results. In the case of correlational data, it is relatively simple to square the coefficient of correlation and secure an estimate of the amount of variance accounted for by the particular set of correlates. With other methods of analysis, the implications may not be as readily apparent, but it is equally important that an estimate of the variance accounted for by the experimental manipulation be provided.

The importance of such analyses can be illustrated from some recently published reports, as well as from reviews of other unpublished material. Particularly when large samples are used, the author should feel obligated to stress the implications of his/her findings in terms of the variance accounted for by the variables under study. For example, in the published report of the current activities and preferences of a sample of 855 clinical psychologists (Garfield & Kurtz, 1976), numerous findings at the .01 level of significance or better were secured. Some may have been the result of the numerous comparisons made, but the sample size appeared to be of some consequence in this regard. Correlations of .10 that were highly significant statistically were consequently considered to be of little practical importance, since they accounted for only 1% of the variance. In another published study of over 1,000 clients in a number of community mental health centers, several correlations of around .10 were reported as highly significant, even though they were obviously of little significance clinically or socially (Sue, McKinney, & Allen, 1976). For example, the correlation

between diagnosis and premature termination was .10, and this was significant at the .001 level of confidence. However, by itself, such a significant finding accounts for a negligible amount of the variance.

In another study, a mean difference of 4 on one subtest of the Wechsler Adult Intelligence Scale (WAIS) was obtained between two different administrations of the test to a sample of schizophrenic patients. This difference was reported as being significant at the .05 level of confidence, and the author went on to offer some detailed conclusions pertaining to this "significant" finding. Results of such magnitude, however, would not appear to be of much practical value.

Apart from going beyond the obtained levels of statistical significance to report the amount of variance accounted for by the variables studied, other analyses may be required if the data are presumed to have some value for clinical diagnosis. The means and standard deviations for the clinical samples studied should be reported, and the extent of overlap of the distributions should also be clearly stated. Of value also for clinical diagnosis are the number of subjects who would be correctly diagnosed or classified by the diagnostic procedures evaluated and those who would be misclassified—that is, data on false positives and false negatives. These are certainly important data for any purportedly effective diagnostic procedures, but they are not always secured or provided in reports of research in this area. It should be apparent that a particular diagnostic technique may differentiate two or more clinical groups at the .05 level of significance but yet produce so many false positives or negatives that it has very limited clinical utility. It would appear incumbent on investigators to analyze their data in terms of such considerations and to present their analyses in a clear fashion.

Base Rates

Another problem in some types of research on clinical diagnosis pertains to the old issue of base rates. Even though this important matter was raised some years ago and is not an unfamiliar topic in either the areas of diagnostic assessment or psychotherapy (Gathercole, 1968; Meehl & Rosen, 1955),

at least a certain number of studies appear to disregard it even when it is clearly relevant. Consequently, although the matter of base rates is not a new issue nor even a complex one, I will devote some space to it and provide a few examples of what is involved.

I can start this brief discussion by referring to an earlier experience of my own. Working as a clinical psychologist in a Veterans Administration hospital 30 years ago, I collected some Rorschach test data on a moderate sample of patients who had been diagnosed as schizophrenic by the clinical staff. I compared my diagnostic impressions based on the test data with the clinical diagnoses on these patients and, among other things, noted that my diagnoses agreed with the staff diagnoses in about 67% of the cases. I prepared a paper on this study and submitted it to the branch chief psychologist, who at that time was David Shakow. He returned it to me with his comments. Among the suggestions he made was that I secure the base rates for diagnoses of schizophrenia in my hospital. I was rather reluctant to comply with his suggestion but did a quick survey of admissions for a limited time and discovered that the number of cases diagnosed as schizophrenic was just about 67%. In other words, my diagnostic work did not exceed the base rate for diagnoses of schizophrenia, and diagnosing every admitted patient as a schizophrenic would have been as accurate as diagnoses derived from my Rorschach examinations.

The matter of base rates, thus, is of some consequence, and it is certainly a factor that must be appraised in evaluating certain kinds of research on clinical diagnosis. Apart from the fact that diagnostic procedures to be effective must clearly exceed the base rates for a particular disorder, attention to such matters along with attention to false positives and negatives indicates the potential difficulties in using diagnostic procedures for disorders in which the incidence is very low, as is the case for suicide. In the latter instance, a diagnostic or predictive measure must be extremely effective in discriminating the specific type of cases being evaluated if they are to be clinically useful.

Thus, for specific kinds of problems per-

taining to differential diagnosis or prediction, it appears essential for the researcher to provide data on the base rates for the disorders of interest and to clearly show the advantages as well as the disadvantages for the procedures being evaluated. Listing the percentage of correct hits or diagnoses obtained by a particular diagnostic technique is not sufficient. Along with other criteria, attention must be paid to the matter of base rates.

Inadequate Data Presentation

The present designation includes a variety of frequently small but, nevertheless, important oversights that are apparent in at least some manuscripts. Among these are such matters as not providing basic information on the measures or scales used, necessary information on important variables, how diagnoses were secured, how the subjects were selected, and similar aspects. The examples to be discussed will not be exhaustive but rather will illustrate the kinds of problems encountered when such information is not included in the report.

In the current era in which a large number of patients, both inpatients and outpatients, are receiving medication of various kinds, it is extremely important that this information be provided in the report of the research. If the drugs or medications used have any potency, they are bound to have some influence on the behavior and mental functioning of the subject studied. Not only must the taking of medication be clearly mentioned, but the medication used, the dosages, and in many cases, the duration of medication should also be specified. If part of a group of subjects is on medication and a part is not on medication, there are clearly problems in mixing the results of these two subgroups and treating them as one relatively homogeneous group. In a related fashion, comparing a clinical group of subjects receiving medication with a control group that is not must influence the kinds of conclusions that can be drawn from such comparisons. If one is attempting to compare the effects of drugs on two groups of comparable patients, then, of course, the previous comparison would be feasible, providing a placebo was used with the control group. However, if the comparison were made

to compare the mental functioning, behavior, or personality characteristics of a given diagnostic group with some other group, then such a comparison would provide results contaminated by the influence of the medication. Although this appears quite obvious, it is surprising how frequently this kind of a problem is ignored or glossed over in manuscripts submitted for publication.

Another problem concerns the lack of adequate information presented on the particular techniques or methods of appraisal used in the research study. This is of particular consequence when such procedures are not standardized ones, are not very well-known, or have been constructed by the investigators for their specific research, often without adequate reliability or validity studies. It is the responsibility of investigators to describe clearly the procedures and techniques that they have used in their study so that others can fully understand what has taken place and be able to attempt possible replications of the study if they so desire.

With the pressures for publication evident today and the corresponding desire to use journal space as efficaciously as possible, it is understandable that editors want manuscripts to be as brief and concise as possible. Unnecessary verbiage should, of course, be deleted from all manuscripts. Nevertheless, this does not mean that significant information about a research project should be omitted. When relatively unknown tests or rating scales are used, for example, the investigator should describe them in sufficient detail so that the reader has a clear understanding of the techniques used and can appraise their suitability for the sample used and the problem under investigation. The investigator should also provide pertinent data concerning the reliability and validity of the procedures used in these instances as well. Since the value of the results secured in any research study is dependent on the types of measures used, such information would appear to be essential.

The Lack of Cross-validation

One of the apparent causes of concern about much past and current research, not limited to clinical diagnosis alone, is the large num-

ber of conflicting findings in the literature and the apparent difficulty encountered in the replication of published findings. This has become such a regular feature of research in clinical psychology that many review articles tend to sum up the number of positive and negative studies on a given topic. If over half of the studies appear favorable to a given proposition, then that view is judged to have some support, even when the studies vary greatly in quality. This is a somewhat risky kind of approach to drawing conclusions about some technique or finding. If a particular treatment, for example, is found helpful in eight investigations and harmful in four, should it be concluded that the overall impression is that the treatment is helpful? The fact that there are such conflicting findings should make us suspect some possible limitations in the research reported and withhold judgment until we are able to explain the confused findings.

Most likely, the discrepancies among research reports are due to subject variables, small samples, and variations in diagnostic procedures, as well as to chance variables. Because of the kinds of problems already mentioned concerning variations in assigning clinical diagnoses, in subject samples, in the kinds of settings used, and sometimes also in the procedures used, the findings from any single investigation can only be viewed as suggestive. Although attempted replications by other investigators in different settings appear to be essential in appraising the value of any investigation, there is one procedure that most researchers could use to improve the reliability of their findings. This is simply to secure enough subjects in their experimental and control groups so that all groups could be randomly divided into two groups. The study is then conducted with one set of groups functioning as the initial group, and the second set can then be used for cross-validation. This is a relatively straightforward procedure that has been used in some studies, but relatively infrequently.

Examples of studies that contained such attempts at cross-validation and that also illustrate their value are Garfield and Wolpin (1963); Lorr, Katz, and Rubinstein (1958); and Sullivan, Miller, and Smelser (1958). In

the study by Sullivan et al., two cross-validations were carried out on MMPI profiles of patients who terminated prematurely from psychotherapy, and the results of the cross-validation attempts were clearly important in the final conclusions drawn. In this study, significant differences between premature terminators and those who continued in therapy were found for several MMPI scales for *each* of the several groups of subjects investigated. However, *none* of the scales that differentiated the groups under study showed a consistent pattern. For each separate appraisal different scales were found to be significant in their differentiation. Consequently, the findings secured in the first appraisal were not supported in the subsequent cross-validations, and the final conclusions reached were different than they would have been if the cross-validations had not been attempted. These authors, at the conclusion of their article, emphasized the necessity of cross-validation, and I strongly concur with their conclusion.

Although attempted replications or cross-validations by other investigators would also be required to fully evaluate the significance and utility of findings reported in a single investigation, the procedure suggested here is still a useful one for increasing the potential value of single studies. Most likely it might require larger initial samples of subjects so that the samples can be divided into two groups for purposes of cross-validation. However, such extra effort and cost would appear to be justified by the possible increase in the significance of the findings secured, and perhaps some premature generalizations and conclusions might be negated.

Because of the complexities and potential problems in conducting research in the area of clinical diagnosis, investigators in a majority of instances should attempt some sort of cross-validation before reaching conclusions and before submitting their research reports for publication. Their findings would tend to have greater reliability and validity, and, hopefully, the parade of contradictory findings that appear in our journals would be decreased. Initial findings of a single investigation that have not been cross-validated but that appear to be of possible importance should be submitted for publication only as

brief reports. As already mentioned, such findings that have not been cross-validated can only be viewed as suggestive.

Miscellaneous Issues

There are also several other issues that can be mentioned somewhat briefly. All investigators can be presumed to be aware of these matters, but at times there appear to be some lapses in attention to them.

One basic consideration that is sometimes neglected concerns the careful checking of results and computations before submitting a manuscript for publication. This seems so obvious as to not require any mention here, but errors in computation do occur. In the apparent haste to get the manuscript ready for submission for publication, an author may fail to detect some rather obvious, and at times important, errors in his or her manuscript. In some cases, the conclusion drawn from the study may have to be greatly modified, which is a decided embarrassment to the author involved.

Related to the above are occasional misuses of statistical procedures. In the case of chi-square, inflated *N*s, based on the number of observations instead of the number of subjects, can provide spurious results. The use of parametric statistics for nonparametric data is another example. One of the more common errors is treating highly related measures or ratings as independent measures. Because a number of us are deficient or not up-to-date on our knowledge of statistical procedures, it is a wise procedure to seek consultation from other more knowledgeable individuals when we are planning studies, when we are ready to analyze our data, and when we are interpreting and writing our research results.

Another matter that is occasionally noted pertains to the interpretation of findings secured by means of correlations. Coefficients of correlation are useful indices of the relationship between two variables. However, the relationship between two variables cannot be used as a basis for judging causality. For example, one may find a significantly high correlation between marital status and length of hospitalization. One might be able to say that marital status is a favorable prognostic in-

indicator for future discharge from the hospital, but one could not say that getting married causes a short stay in the hospital. It is conceivable that better integrated individuals or less asocial ones are more likely to get married and that these variables or others like them may possibly be related to outcome.

Another practice that I have occasionally noted is that of drawing major conclusions or inferences from very scanty or minor bits of data. One should not let his or her enthusiasm exceed the quantity and quality of the actual data secured. In one of the studies mentioned earlier, for example, in which a mean difference of .4 on a subtest of the WAIS was secured between two different administrations of the test, the author offered some detailed (and speculative) reasons to account for this "significant" result. Results of such magnitude do not appear to require any lengthy interpretations of the data.

Clinical Diagnosis: Some Concluding Comments

Because the issue of clinical diagnosis is such a complex, controversial, and yet crucial issue, a few final words on this topic are in order. It is readily apparent that any research on diagnostic groups, for example, schizophrenia or manic-depressive psychosis, can be no better than the validity or meaningfulness of the diagnoses secured and used. This problem has plagued research in this area for many years. On the one hand, systematic and reliable classification of subjects facilitates research and the accumulation of potentially meaningful data about types of psychopathology. On the other hand, if the classification scheme used for such research is beset with problems of clarity, reliability, and validity, the results based on such classification are bound to be limited in their usefulness. An unreliable and loose scheme is bound to produce unreliable and variable results. The most highly quantified data and the most exacting statistical analyses cannot provide worthwhile conclusions if the assumptions or foundation on which they are based are weak.

If research based on clinical diagnosis is to have any theoretical or clinical value, more

attention will have to be paid to improving the selection and specification of research samples. Otherwise, we will continue to have the confusing array of conflicting and contradictory findings that makes its way into the published literature each year, despite the high rejection rate of journals published by the American Psychological Association.

Although it will not be easy to improve the situation greatly, it is my belief that some improvements can be made in the current situation. Editors should openly state that articles based on broad diagnostic categories will usually not be accepted for publication. That is, reports on "schizophrenia," "brain damage," and the like, will be considered too diffuse to merit publication. Instead, greater specificity will be required in the future. In line with the suggestions discussed earlier in this article, the authors of studies will be expected to specify and describe their research subjects in ways that give more meaning to the diagnostic categories studied. In cases diagnosed as schizophrenic, for example, the patients will have to be described also in terms of measures such as process-reactive; good or poor premorbid; the MMPI; IQ; factored scales such as those developed by Lorr, Klett, McNair, and Lasky (1962); and the like. Major symptoms of the subjects should also be described and rated on reliable scales. To treat schizophrenia as a single meaningful category or disorder is to court chaos and disaster.

In a similar manner, brain damage has for too long been treated as if it were a unitary disorder when actually a great variety of disorders and impairments have been subsumed under this crude designation. The site of the disorder in the brain, the type of lesion, the type of onset, and the source of the disorder or injury are all pertinent. Even though some progress in the direction suggested here has been made in recent years, much more needs to be done along these lines. This is particularly true as new and equally vague categories such as the hyperactive child, minimal brain damage, infantile autism, and learning disorders begin to receive both popular and research attention.

Unless we pay attention to the research issues involved and try for as much specifica-

tion as possible of the disorders we are studying, our research efforts will only produce conflicting and disappointing results, and clinical practice will be the loser. We can learn from our past deficiencies and, hopefully, strive to improve the quality of our investigative efforts.

References

- American Psychiatric Association. *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, D.C., Author, 1968.
- Beck, A. T. *Depression: Causes and treatment*. Philadelphia: University of Pennsylvania Press, 1972.
- Finlayson, M. A. J., Johnson, K. A., & Reitan, R. M. Relationship of level of education to neuropsychological measures in brain-damaged and non-brain-damaged adults. *Journal of Consulting and Clinical Psychology*, 1977, 45, 536-542.
- Garfield, S. L. An evaluation of Wechsler-Bellevue scatter patterns in schizophrenia. *Journal of Consulting Psychology*, 1949, 13, 279-287.
- Garfield, S. L. *Clinical psychology: The study of personality and behavior*. Chicago: Aldine, 1974.
- Garfield, S. L., & Kurtz, R. Clinical psychologists in the 1970s. *American Psychologist*, 1976, 31, 1-9.
- Garfield, S. L., & Sundland, D. M. Prognostic scales in schizophrenia. *Journal of Consulting Psychology*, 1966, 30, 18-24.
- Garfield, S. L., & Wolpin, M. Expectations regarding psychotherapy. *Journal of Nervous and Mental Disease*, 1963, 137, 353-362.
- Gathercole, C. E. *Assessment in clinical psychology*. Baltimore, Md.: Penguin Books, 1968.
- Hamilton, M. A. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, 1960, 23, 56-61.
- Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.
- Lick, J. Statistical vs. clinical significance in research on the outcome of psychotherapy. *International Journal of Mental Health*, 1973, 22, 26-37.
- Lorr, M., Katz, M. M., & Rubinstein, E. A. The prediction of length of stay in psychotherapy. *Journal of Consulting Psychology*, 1958, 22, 321-327.
- Lorr, M., Klett, J., McNair, D. M., & Lasky, J. J. *Manual for the Inpatient Multidimensional Psychiatric Scale*. Palo, Alto, Calif.: Consulting Psychologists Press, 1962.
- Meehl, P. E. High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology*, 1971, 77, 143-148.
- Meehl, P. E., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 1955, 52, 194-216.
- Pasamanick, B., Dinitz, S., & Lefton, M. Psychiatric orientation and its relation to diagnosis and treatment in a mental hospital. *American Journal of Psychiatry*, 1959, 116, 127-132.
- Sue, S., McKinney, H., & Allen, D. Predictors of the duration of therapy for clients in the community mental health system. *Community Mental Health Journal*, 1976, 12, 365-375.
- Sullivan, P. L., Miller, C., & Smelser, W. Factors in length of stay and progress in psychotherapy. *Journal of Consulting Psychology*, 1958, 22, 1-9.

Received October 20, 1977 ■

Methodological Considerations in Clinical Neuropsychological Research

Oscar A. Parsons and George P. Prigatano
Department of Psychiatry and Behavioral Sciences
University of Oklahoma Health Sciences Center

Methodological problems and issues in clinical neuropsychological research are discussed for four types of neuropsychological studies: (a) differential diagnosis, (b) basic brain-behavior relationships, (c) effects of noxious agents or factors on brain-behavior relationships, and (d) rehabilitation of neuropsychological deficits. Recommendations as to how to handle these methodological problems are made. The characteristics of good case study reports are presented.

Neuropsychology is the study of brain-behavior relationships; clinical neuropsychology is the application of empirically established facts concerning these relationships, and theories derived from them, to clinical problems. During the last decade we have witnessed an outpouring of research in clinical neuropsychology. All indications point to a continuation of this healthy and vigorous area of research. In this article we discuss some of the methodological problems that, if carefully considered and acted on, could substantially improve the quality of the research. (And we include ourselves as recipients of that admonition!) In some instances the methodological points made will have much in common with any psychological experiment; in other instances problems that are fairly unique to neuropsychological research will be addressed.

Before consideration of specific methodological problems, it is appropriate to point

out that competent clinical neuropsychological research demands mastery of several areas. The traditional areas of measurement of human behavior (psychometrics) and psychopathology, coupled with a serious study of the central nervous system (e.g., neuroanatomy, neurophysiology, and neuropathology) and exposure to clinical neurological concepts and procedures, are necessary, in our opinion, for the level of sophistication demanded by this complex field. This is not to say that one cannot conduct meaningful neuropsychological research without knowing the microscopic anatomy of the brain; we are saying that good neuropsychological research demands a background often lacking in psychologists' training. This is a remediable condition.

Given an adequate background, a second factor conducive to good research in this area is the continual recognition that behavior is the final common expression of a tremendous number of possible factors. We emphasize this somewhat trite and elementary maxim because it is seemingly frequently ignored. For example, cognitive "deficits" are characteristically found in brain-damaged patients, but they are also present in patients with schizophrenia (G. Goldstein & Halperin, 1977), depression (Miller, 1975), anxiety (Chapman & Wolff, 1959), or culturally deprived or educationally disadvantaged persons (Amante, VanHouten, Grieve, Bader, & Margules, 1977). In the Halstead-Reitan Battery (Reitan & Davison, 1974), much attention is

Preparation of this article was supported in part by U.S. Public Health Service, National Institute for Alcohol Abuse and Alcoholism Grant AA 01464 to the first author.

George Prigatano is a visiting scholar at the Department of Psychology, Stanford University, Palo Alto, California.

Requests for reprints should be sent to Oscar A. Parsons, Department of Psychiatry and Behavioral Sciences, University of Oklahoma, Health Sciences Center, P.O. Box 26901, Oklahoma City, Oklahoma 73190.

given to differential left-hand and right-hand motor performance as an indication of laterality of cortical damage. However, poor motor performance with either hand may result from peripheral motor and sensory injury of a subtle nature. Unless this possibility is thoroughly explored by careful history taking, erroneous inferences could be made about the brain state. Our point here is that if the neuropsychological researcher keeps this maxim in mind, possible alternative explanations or confounding variables may be anticipated, thus resulting in a more acceptable research report.

Let us consider, for example, five independent variables that most experimenters would agree are of importance in behavioral research: age, education, sex, socioeconomic level of subjects, and experimenter or examiner characteristics. Are these variables also important in neuropsychological research?

Age

Heaton, Baade, and Johnson (1978) have recently reviewed 94 studies in which neuropsychological test scores of psychiatric patients and brain-damaged patients were compared. In 38% of these studies, age was either not mentioned or the groups differed with respect to age. Given the well-known relationship of age to cognitive, perceptual, sensory and perceptual-motor, and information-processing variables, most of which are being measured in neuropsychological studies, it is vital that age receive close scrutiny. The safest procedure is to equate groups being studied on age both as to means and standard deviations. If age is significantly different for groups, and age matching by dropping or adding subjects is not possible, there are several alternatives. First, it should be determined if the dependent variable is in fact correlated with age. If it is not correlated in the brain-damaged, control, or other experimental groups, then it is likely that age could be discounted as a significant factor.

Note, however, that a lack of correlation in the brain-damaged group could not be used to disregard the influence of age. Heterogeneity of damage to the brain could obfuscate any general relationship. For example, Prigatano and Parsons (1976), in a cross-validated

study, found that the Category and Tactual Performance Test measures on the Halstead-Reitan Battery correlated significantly with age in both brain-damaged and non-brain-damaged groups but that the Rhythm, Speech Perception, and Tapping tests correlated significantly only in the non-brain-damaged groups. Apparently, the brain-damage effect was sufficient to override the effects of age in the latter variable. Incidentally, the correlations of age with the Halstead Impairment Index in the non-brain-damaged groups were .57 and .64; in the brain-damaged group they were, respectively, .33 and .44. All correlations were significant, but they were consistently lower in the brain-damaged group.

Another method of handling age difference among groups, once the relationship between age and dependent variables has been established, is to adjust the dependent variable scores by analysis of covariance and then test for differences. Partial correlational techniques could also be used. To repeat, however, the most clear-cut solution is to equate groups on age means and variability.

Education

Heaton et al. (1978) were "reassured" by the 60% of the studies that they reviewed, which stated that the groups under investigation had comparable educational background. What about the 40% who did not? Does education make a difference? Two recent articles have addressed this question. Prigatano and Parsons (1976) correlated education and Halstead Battery performances in brain-damaged and other patient groups. Neither the brain-damaged nor the psychiatric group had significant correlations. However, in a group of non-brain-damaged general medical-surgical patients ($N = 50$), they found significant correlations with six Halstead measures. It seems likely that psychiatric disorder and brain damage introduce enough variance to reduce the correlation found in a nonpsychopathological population. The effects of education were demonstrated in a clear-cut fashion in a recent study by Finlayson, Johnson, and Reitan (1977). These investigators compared brain-damaged and control subjects, stratified in three levels

of education, on the Halstead Neuropsychological Battery. Both level of education and brain damage had a "pronounced effect" on scores; the lower education groups scored lower on the Halstead measures, as, of course, did the brain-damaged group. Our recommendations to investigators as regards the possible effects of education are the same as we discussed for age.

Sex

Given the popularity of this variable in most dimensions of our lives, it seems rather strange that the possible differences between males and females in neuropsychological studies has received little attention. In fact, in the Heaton et al. (1978) article, sex is not even mentioned, let alone introduced as an important methodological variable! Yet there have been at least two neuropsychological studies that have found that the sex variable is of some importance. Lansdell (1968) studied male and female epileptic patients who had temporal lobe tissue removed. Difference in interactions of hemisphere and sex on perceptual variables suggested that subtle differences may exist in male and female brains as regards such variables. McGilone and Kertesz (1973) concluded that brain damage in males leads to greater functional asymmetry, especially in visual-spatial analysis (Block Design), than in females. In children, Witelson (1976) has demonstrated clear-cut male-female differences in tactile-spatial performance, with right-hemisphere-governed performance significantly better in males than in females.

Our suggestion as regards this variable is that when possible, groups should be equated on number of males and females; if this is not possible, then sufficient numbers of either ~~male~~ or females should be included to analyze the data for at least ~~one~~ sex with a respectable *N*. If small numbers of either sex are included in the total *N*, it should be demonstrated that obtained differences in dependent variables are similar to those effects obtained with the different-sex larger *N*.

Socioeconomic Level

This variable is closely related to education, and so in most adult studies, education is

sufficient. However, as education levels continue to rise and "social" promotions are the rule rather than the exception, it may be important to again consider the broader variable of socioeconomic level. Another aspect is the occupation of the subjects. Has there been overlearning of specific skills, which in turn could lead to spurious results? For example, manual workers and bookkeeper clerks might well differ in patterns of strength in perceptual-motor and verbal-calculation skills. Depending on the title of occupational representation in a given sample of brain-damaged patients, many different inferences can be made. The socioeconomic variable is quite important in neuropsychological investigations of children. Indeed, in a recent study, Amante et al. (1977) concluded that levels of neurological integrity vary along a socioeconomic gradient. The relevant factors associated with social causation include malnutrition, reduced environmental stimulation, and inadequate obstetrical and pediatric care.

We recommend that adult studies in neuropsychology consider the occupational variable along with education. Certainly with children, the general family socioeconomic background should be specified, as educational levels will be roughly equivalent for groups through age 16.

Examiner Characteristics

Another independent variable to be considered is that of the experimenter, or more specifically the examiner or tester. Schafer (1954) some years ago eloquently described the various problems faced by the psychological test examiner. Among them was the "pressure" to obtain a "score," a pressure that can often lead to an inattention to the patient's interpersonal needs during the examination or even to poor test administration. Inexperienced or poorly trained examiners making such mistakes may well fail to elicit maximum effort or motivation on the part of the patient. Without the latter, the validity of tests such as the Halstead-Reitan Battery may be questioned (Reitan, Note 1). Studies in clinical neuropsychology seldom describe the degree of expertise of the examiners. Certainly there are varying degrees of sophistica-

tion of the neuropsychology technicians used in the various clinical and research laboratories throughout the country. Yet, clinical neuropsychological research is often published with no mention of how this variable may have influenced their reported findings.

The social psychological research of Rosenthal (1968) has certainly pointed to the possible impact that experimenter or examiner expectations may have on subjects' performance or the evaluation of subjects' performance. Schachter (1964) has argued effectively that the perceived emotional environment can have a major influence on a person's subjective emotional state and reactions, which in turn can affect performance. Although there are relatively few neuropsychological studies that address these questions, Parsons and Stewart (1966) found that brain-damaged patients showed less improvement on a perceptual-motor neuropsychological test when examined by a "disinterested-factual" examiner as compared to one who was "supportive" and interested in reducing the patient's test anxiety.

The demonstration that the interpersonal climate is important in examining brain-damaged patients is important in view of the growing use of technicians and computer-run laboratories. A disinterested examiner or an interaction with a computer might create the type of atmosphere that interferes with the maximum level of performance in some patients or groups of patients but not in others, leading to conflicting results across laboratories.

We recommend that neuropsychological research reports state the level of sophistication and training of the persons who conduct the examination. A short description of the interpersonal climate would also be helpful (e.g., "patients were encouraged frequently"; "if tired, patients took a short break"; "instructions were given clearly, and testing did not proceed until the patient demonstrated understanding"; "supportive reassurance was appropriately given"; etc.).

Approaches to Neuropsychological Investigations

Neuropsychological studies customarily have fallen into three major types. In the

first type, the objective is the differential diagnosis of brain-damaged patients from non-brain-damaged control and other psychopathological groups, an applied or clinical problem. The second type seeks to identify and describe the general and specific changes in brain-behavior relationships as a function of brain states. The intent of these latter studies is usually to contribute to our basic understanding of neuropsychological relationships. The third type consists of studies that attempt to ascertain the effects of certain noxious agents (e.g., metabolic disease, drugs, trauma, etc.) on the brain-behavior relationships. Looming on the horizon, however, is a fourth type of studies, the new and challenging field of rehabilitation of neuropsychological deficit behavior. Of course, there are many studies in which several objectives are present, but for our present purposes we will consider the four types separately. Finally, we shall present our notions of what constitutes a good case study.

Differential Diagnosis

A problem area that has preoccupied psychologists since the rise of clinical psychology in the late 1940s is that of identification of brain-damaged patients. The objective of these studies is to develop behavioral measures that will discriminate a group of patients typically identified as "brain damaged" from patients who have another psychopathological disturbance (e.g., schizophrenia, depression, or anxiety) and from non-brain-damaged, nonemotionally disturbed control subjects. Although there are many investigators who have wondered whether continued efforts in this area are warranted (Parsons, 1970; Smith, 1975; Spreen & Benton, 1965), the continued interest and vigorous pursuit of such studies suggest that the working clinicians believe otherwise.

Subject Characteristics

The primary methodological problem is that of subject selection and subject characteristics. Let us consider the brain-damaged patients. It is not unusual in neuropsychological studies to find that the patients were de-

clared "clearly brain damaged" from the medical records by unknown or at least unidentified physicians or other persons (Heaton et al., 1978). There seems to be an assumption that perfectly reliable criteria were used. It should come as no surprise, however, to learn that even neurologists do not always agree on diagnosis of brain damage. In one of the classic studies (Fisher & Gonda, 1955), two skilled neurologists made a judgment as to whether damage existed rostral to the foramen magnum in patients who were given thorough neurological and other biomedical tests. This judgment was based on review of the records and included clinic follow-up reports and occasional autopsy reports. The agreement was 86% under these superior diagnostic conditions. Surely under less favorable conditions the percentage of agreement would drop.

Given the problems of cost, availability, and time, the preferred method of having two neurologists independently examine the patient probably cannot be achieved by most neuropsychological researchers. One solution would be to have at least one neurologist review all available data. Criteria of positive neurological examination plus at least one positive biomedical test (e.g., computerized axial tomography [CAT] scan, skull films, arteriogram, pneumoencephalogram, electroencephalogram, and so forth) could be used. A selected sample could be blindly rated again by the neurologist to give some indication of the reliability of his or her own diagnostic statements. The point is that more attention must be given to the defining criteria of brain damage, especially the qualifications and reliability of the clinical judges, the latter being preferably the relevant medical specialists (Jacobs, 1977). After all, the behavioral measures can only be as effective in diagnostic statements as the reliability and validity of the criterion measure of brain damage permit.

As Piotrowski (1940) pointed out almost 3 decades ago, one of the best ways to determine the diagnostic utility of neuropsychological tests would be to apply the tests to a sample of persons referred for suspected brain damage. Some of these people would show brain damage; others would not, the latter

constituting what has more recently been termed *pseudo-neurologic* (Matthews, Shaw & Kløve, 1966). If the neuropsychological tests (independent of the diagnostic process) can distinguish brain-damaged from non-brain-damaged persons, the tests have a diagnostic potential. Note that this is a more exacting procedure than the method mentioned earlier (i.e., using clearly diagnosed patients) and much more difficult to achieve practically. As neuropsychological testing has become more widespread and of demonstrated effectiveness (Filskov & Goldstein, 1974; G. Goldstein, 1974; Reitan & Davison, 1974; Satz, Fennell, & Reilly, 1970; Smith, 1975), it undoubtedly influences decisions about diagnoses. In retrospect these influences are quite difficult to partial out. Further, a non-invasive superior biomedical test, the CAT scan, is now in use at most centers. Patients who previously presented diagnostic problems, because of the physicians' understandable reluctance to use invasive techniques, are now diagnosed earlier in the process. Parenthetically, use of the CAT scan as an important criterion measure in neuropsychological studies will be routine in the future. It is important to note, however, that reliability studies of CAT scans are only now appearing and that normal values and ranges for given laboratories are frequently inadequate. At the present time, users of the CAT scan in neuropsychological studies would be advised to quantify their measurements as much as possible and to present interrater reliabilities.

Base Rates

To return to the problem of subject characteristics in studies of differential diagnosis, the ultimate criterion of a good diagnostic test is the "hit rate." The more true positives (brain damaged) and true negatives (non brain damaged) and the less false positives (not brain damaged called brain damaged) and false negatives (brain damaged called non brain damaged), the better the test. However, interpretation of the efficacy of hit rates is highly dependent on the base rates in the populations sampled. Contrast, for example, the high incidence of brain-damaged patients on the neurology service with the typically

low incidence seen on psychiatric services in general hospitals. The cost efficiency of neuropsychological tests (Rimm, 1963) will differ under these conditions. Clinical neuropsychological researchers should be thoroughly familiar with the base-rate problem (Gordon, 1977; Heaton et al., 1978; Krug, 1971; Rimm, 1963; Satz, 1966).

Test Variables

What should guide the researcher in test selection in studies of differential diagnosis? First, researchers should be aware of the studies that have compared hit rates with single and multiple tests. Spreen and Benton (1965), in their review, reported that individual neuropsychological tests discriminated brain-damaged from normal controls with an average hit rate of 71%; with several tests the cumulative predictive value was 80%. They concluded that the search for screening devices had reached its culmination point and served its purpose. Heaton et al. (1978) reported a median hit rate of 75% for 84 classification attempts involving a variety of psychiatric disorders but omitting process or chronic schizophrenia. Single tests and combinations of tests appear to have similar hit rates. The Halstead Impairment Index, an index based on seven tests, frequently gives the same level of hit rate as individual component tests (Smith, 1975).

In other words, for differential diagnosis there seems little warrant for constructing a large battery of tests. This is not to say that a large sampling of test behaviors is not important, rather it means that to answer the specific question of differential diagnosis, most good single tests give practically the same level of discrimination as combinations of tests. Of course, for gaining information, beyond that of diagnosis per se, enlarging the range of behaviors studied (i.e., multiple tests) is highly desirable. In fact, in most clinical settings, the emphasis today is on specifying the nature and location of the brain damage. As the field moves toward a greater emphasis on rehabilitation, as we will discuss later, a more extensive sampling of behavior may be required.

Reliability of tests is always a concern for

the researcher. What do we know about the reliability of neuropsychological tests? Matarazzo, Matarazzo, Wiens, and Gallo (1976) presented test-retest data on the Halstead-Reitan Impairment Index for four groups of subjects: normals, schizophrenics, and two samples of organic patients. In the patient groups the test-retest reliability was reassuringly high (.83, .63, and .82), considering that test-retests in clinical populations are always attenuated by change in patients' conditions over time. In the normals it was low (.08), but this was due to the latter's extremely low Impairment Index scores; actually 29 out of 29 were classified as normal in each testing, for 100% accuracy. Of course, if the investigator is devising a new test, test-retest reliability coefficients should be presented. If for some reason this is not possible, other estimates of reliability such as split-half correlations can be used. As we will discuss later, reliability of the test is significant in specification of group difference in performance. We are past the day when investigators try out a new test in brain-damaged and control groups, find large differences, and optimistically submit the results for publication without considering such basics as reliability.

Statistical Analysis

In recent years the advent of computer analyses has enabled multivariate statistics to be applied to problems in differential diagnoses. These analyses have the advantage of maximizing differences among groups and identifying the measures on which groups differ most (G. Goldstein & Halperin, 1977; Swiercinsky & Warnock, 1977). It should be recognized that the sensitivity of multivariate analyses in accentuating group differences capitalizes on the chance variance that may be present as much as on true variance differences. Any study that purports to have heightened discrimination through multivariate analysis should be cross-validated. For example, in discussion of the use of stepwise regression multivariate analysis, Cohen and Cohen (1975) advise that if the results are "to be substantively interpreted, a cross-validation of the stepwise analysis in a new sample be undertaken, and only those conclusions

which hold for both samples should be drawn" (p. 104).

In our opinion the search for the Holy Grail (i.e., the neuropsychological test(s) that will give a high discrimination of brain-damaged patients from controls or other psychopathological groups in all types of settings, with all socioeconomic levels) is not likely to be as productive as other research approaches. If one feels compelled to continue this search, then we suggest that the investigator follow these basic suggestions: (a) Give specific information on the reliability of the diagnoses (done preferably by relevant specialists) for all of the patient groups studied; (b) give some estimate of the base rate of brain damage in the samples studied and compare hit rates with them; (c) give some indication of the reliability of the tests used for measurement; and (d) conduct a cross-validation study to determine whether the statistically maximized hit rates will hold for another sample.

Studies Elucidating Basic Brain-Behavior Relationships

In these studies the concern is with gaining basic understanding of the qualitative differences in brain-behavior relationships. What are the functions of the right and left hemispheres? Are there different behavioral patterns characteristic of localized lesions? These and many other questions are answered by studying patients who are identified on the basis of known damage to the brain in certain areas (e.g., left-hemisphere lesion vs. right-hemisphere lesion patients). The patients are given neuropsychological tests and, if differential results are obtained, inferences are made about the role of those affected neuro-anatomical regions in behavior.

General Versus Specific Effects

At the outset it is important to recognize that there are both general and specific effects of brain damage (Adams, 1969; Chapman & Wolff, 1959; Parsons, 1970; Smith, 1975) and that both effects are dependent on the size or extent of the lesion. Chapman and Wolff (1959), in a classic article that should

be read and reread by researchers in clinical neuropsychology, have provided convincing evidence of the relationship; the larger the amount of tissue removed from the brain, the greater the postoperative impairment of the patient. Specific effects on neuropsychological tests will always be embedded in the larger context of these general effects; comparisons of different lesion locations should be made with patients who have roughly equivalent amounts of dysfunctional or destroyed tissue. If patients are being compared who have not had neurosurgical intervention, this may be a difficult condition to meet. However, some estimate could be made by neurologists' ratings of severity of the effects of the damage (Parsons, Vega, & Burn, 1969).

Variables Related to the Changed Brain

There are a number of aspects of the lesion or disease process to be considered. One variable of utmost importance is duration of the lesion or time since insult to the brain. Smith (1975) provided a clear discussion on the effects of diaschisis (i.e., the effects of localized lesions on more distant parts of the brain). He cited studies indicating disturbed cerebral hemisphere blood flow and cerebral metabolism in both the diseased and healthy hemispheres of patients. Diaschisis is related to the "general" effects noted above but may be separable in that it contributes to the general effects in the period immediately following the insult but diminishes in effect over time.

Another aspect of the duration of lesion variable is the age at which the insult occurs. It is well-known that effects of brain damage as measured in adulthood are dependent on whether the damage was incurred as a child, while the brain was still developing, or after physical maturity of the brain has been reached, that is, by age 15 or 16. (See various chapters in Reitan & Davison, 1974.) Even in adulthood it would appear that duration of several years is a critical value. Over time, static lesions lead to less discriminating differences than lesions that are relatively acute (Reitan, 1966, 1974). Rapidly growing or worsening lesions are also likely to lead to much greater deficits than static lesions even

if the duration is equivalent (Fitzhugh, Fitzhugh, & Reitan, 1961; Reitan, 1966). Combining brain-damaged subjects who may differ in terms of age of onset, time since insult, static or changing nature of lesion introduces diverse sources of variance. Either inaccurate attenuation or accentuation of specific brain-behavior relationships could occur.

The nature of the disease process or dysfunctional brain state is important. Reitan (1966, 1974) has pointed out that cerebrovascular disease, neoplasms, trauma, and degenerative diseases give rise to different patterns of neuropsychological deficits. Different percentages of such patients in any specific localization studies could give rise to misleading generalizations of conclusions regarding specific brain-behavior relationships.

Handedness, Aphasia, Visual Field Defects, Emotional Responses

Four other subject variables deserve noting: handedness, presence of aphasia or visual field defects, and the patients' emotional responses. It is customary to specify the handedness of the population being studied. (Eyedness and footedness are also frequently noted, but their relationship to performance is less clear.) The difference in lateralization effects as a function of handedness is currently the subject of many investigations (See any recent issues of *Cortex* and *Neuropsychologia*.) The relationship is complex and not at all clearly understood, but one conclusion is obvious: Groups to be compared should be composed of similar percentages of right-handed and left-handed subjects.

The presence of aphasia is important in studies of lateralization or localization. Disturbed language functioning may grossly affect the understanding of oral or written instructions or the communications of answers to questions or both (Zangwill, 1969). Inferences made about disturbed, nonverbal, higher cortical functioning in such patients may be incorrectly made. It is of importance to identify the aphasic subjects and the nature of the aphasic disturbances in studies of general effects of brain damage (in which estimates are made of overall levels of intellectual functioning), as well as in lateralization or localiza-

tion studies. A similar caveat can be made for visual perceptual studies in which patients with visual field deficits are used. Faglioni, Scotti, and Spinnler (1969) have presented convincing evidence of the clarification of research findings by considering this variable.

Finally, we should consider the role of the patient's reaction to his or her changed brain state. Parsons (1970) has discussed this issue rather thoroughly, as have other workers (Chapman & Wolff, 1959; K. Goldstein, 1959; Reitan & Davison, 1974; Smith, 1975). Although the typical emotional response of brain-damaged patients to their condition is depression (Parsons, 1970), there are a variety of other effects as described by Chapman and Wolff (1959), including emotional and physical withdrawal, circumscribed interests, increase in premorbid defenses, and lowered frustration tolerance. Kurt Goldstein's (1959) discussion of the organismic changes of the brain-damaged patient places this problem in an even broader holistic framework. Complicating the picture, however, is the recent evidence (Galín, 1974), which suggests that left and right hemispheres may be differentially salient in the control of dysphoric and depressive versus euphoric and sense of well-being feeling states.

In studies that attempt to uncover basic brain-behavior relationships, care should be taken to distinguish general from specific effects. Unless controlled or otherwise accounted for, a number of variables can influence the results and hence the interpretation of specific brain-behavior findings. These variables include age at which the damage occurred, time since onset of lesion, static versus changing (worsening) condition, type of disease or injury, handedness, presence of aphasia and visual-field deficits, and the patient's emotional responses. We suggest that these variables be assessed or otherwise noted in research reports.

Studies of the Effects of Noxious Agents or Factors on the Brain

In this class of studies, the experimenter is usually concerned with the question, what is the effect of "X" (e.g., anoxia, hypoglycemia, LSD, alcohol, liver disease, periods of uncon-

sciousness from trauma, penetrating head wounds, unilateral electroconvulsive therapy, etc.) on functioning of the brain? Whether the variable is an acute study, such as acute alcohol ingestion, or a chronic study as in the neuropsychological changes in chronic alcoholics, the intent of these studies is usually to specify the brain-behavior relationships affected so that appropriate prevention and remediation techniques could be used. Both general and specific effects are likely to be investigated. Most if not all of the methodological variables previously described are pertinent here. Among the specific effects, a popular current line of research is the differential effects of various agents on hemispheric laterality of function. Parsons (1975), for example, has suggested that both the acute and chronic effects of alcohol are more detectable in behaviors associated with right-hemisphere saliency as opposed to behaviors under control of the left hemisphere. If we are interested in whether variable X does result in differential hemispheric test performance, it is advantageous to have tests that differ in content (e.g., verbal and visual paired associates) but are of the same general level of reliability and item difficulty.

Chapman and Chapman (1975) have shown how the more reliable test will lead to the greatest differences among the groups. For example, if the verbal paired associates are of greater reliability than the visual paired associates, even though both might be depressed by variable X, the more reliable test will give rise to the greater and more significant differences. Further, they point out, item difficulty also helps determine the mean difference between groups on tests. Item difficulty affects task variance; the larger the task variance, the larger the separation of groups of high and low ability on the test.

In other words, to examine the effect of variable X on tests measuring right- and left-hemisphere functioning or other comparisons such as frontal versus postfrontal functioning, it would be highly desirable to use tests that are equivalent in reliability and item difficulty. The Wechsler Adult Intelligence Scale (WAIS) satisfies, at least in part, these requirements, which may be one of the reasons why the WAIS continues to be widely

used in neuropsychological research even though it was designed for a different purpose. Although we cannot hope to achieve the comparable psychometric status of the WAIS for most of our neuropsychological tests without national and federally supported effort, we believe that the kind of careful analysis of the psychometric properties of the tests (reliability and item difficulty) as advocated by Chapman and Chapman (1975) provides a rather clear direction for more sophisticated research in this area.

Rehabilitation of Patients with Neuropsychological Deficits and the Case Study Method

Clinical neuropsychologists are becoming more aware of their potential contribution to the field in identifying behavior changes following brain injury and developing rehabilitation programs to improve neuropsychological functioning (Diller, 1976; Lewinsohn, 1973; Diller et al., Note 2). Strategies for improving body image perception, adapting to visual field deficits (Diller et al., Note 2), and improving memory functioning (Lewinsohn, Danaher, & Kikel, 1977) have appeared, which provide excellent examples of what can be done when specific cognitive and perceptual retraining programs are attempted with brain-injured patients. Moreover, there is therapeutic value in providing specific information concerning the nature of a patient's deficits when consulting with patients, their families, and employers. A recent article by Fuld and Fisher (1977) makes this and other important points regarding the management of children with closed head injury. In general, the methodological problems in this area are similar to those discussed previously and will not be repeated. However, as in all cases of remediation, the major problem is identifying the factor(s) at work in change in behavior over time. Without adequate control groups it is difficult to describe causal effects due to the training procedures used. A second problem will be to develop repeatable techniques in which effects of practice can be distinguished from improvement. Our prediction is that this area will gradually become one of the major thrusts of neuropsychology.

Case Study Method

Although group studies in this area are certainly desirable, the time is ripe for the appearance of systematic individual case studies that report on the patterns of neuropsychological recovery following various brain injuries. (For an example of this, see Prigatano, Note 3.) Also, case studies that attempt to assess the efficacy of various treatment or rehabilitative efforts on improving rate and level of recovery are needed. The possible experimental designs that might be used in the single case study are described in detail by Barlow and Hersen (1973). A good example of how individual case studies can lead the way to more systematic group research is seen in the work of Lewinsohn and his associates. Lewinsohn et al. (Note 4) described in some detail a model for assessing memory deficits and their treatment in individual patients. Later, after several patients had been evaluated and treated, group analysis of the data was made, and some important points regarding the value of visual imagery in working with these patients was documented (Lewinsohn et al., 1977).

Also, the use of the case study method to convey the human struggle involved in the recovery process can be a powerful source of inspiration to other workers in the field. Luria's (1972) description of one man's lifetime determination to overcome neuropsychological deficits secondary to a left-temporal-parietal injury is an account that few readers will forget. It serves as an important reminder of what can be done when human will and ingenuity are combined.

Finally, the value of the case study method in making theoretical points is no better demonstrated in the famous case of H.M., in which Milner, Corkin, and Teuber (1968) have shown that specific long-lasting deficits in memory can take place with no concomitant impairment in other global cognitive functions (i.e., WAIS IQ scores). Other studies have also used the case study method to make important theoretical as well as practical clinical observations (e.g., Geschwind & Kaplan, 1962; Wechsler, 1977).

With this background, it may be of some value to review what goes into a methodologi-

cally sound case study and the advantages and disadvantages of this approach. A methodologically sound individual case study clearly defines why the in-depth analysis of a given patient's behavior is generally important to the field. Without this, it is simply an explication of behavior that may or may not have been seen by other investigators and adds little scientific knowledge. Next, the case study must present a clear statement of the patient's history, symptoms, and laboratory data to allow the reader to come to some conclusions about the patient's situation. If this cannot be agreed on, whatever point is being made by the case study will inevitably be obscured. Third, reliable and validated measures, repeated observations, and/or experimental designs devised for individual case studies (see Barlow & Hersen, 1973) should be used whenever possible. This allows the reader to objectively evaluate *data*, not clinical impressions, about a patient.

The advantages of the case study method are that it (a) provides a detailed look at several interacting variables, which is often impossible within the confines of group studies; (b) allows for the appraisal of clinical as opposed to the purely "statistical" significance of a phenomenon; (c) shows documentation of clinical material, which has theoretical importance; and (d) serves as a reminder to investigators that the field must eventually include some explanation of the individual's behavior. The disadvantages are (a) the potential unreliability of a single case and (b) the inability to quantify the effects of potentially relevant variables. With these considerations in mind, the use of the individual case study method can be a useful adjunctive method in the field of clinical neuropsychology. We hope to see more of such studies in the future.

Conclusion

Methodological problems in clinical neuropsychological research are similar in some respects to those encountered in all of behavioral research. Among these are subject variables such as age, education, sex and socioeconomic level, and experimenter characteristics, such as degree of training and attitude. Dif-

ferent types of neuropsychological investigations pose different methodological problems while sharing certain common concerns. Studies of differential diagnosis require attention to reliability of diagnosis, statements as to hit rates in relation to base rates, and reliability of tests. Cross-validation of new findings, especially if multivariate analyses are used, are almost mandatory. Studies that attempt to elucidate basic brain-behavior relationships should be concerned with distinguishing general from specific effects, amount of dysfunctional tissue, duration of lesion, age of onset, type of disorder, handedness, presence of aphasic or visual field defects, and the patients' emotional reaction. Investigations of the effects of noxious agents on brain function, especially differential effects that depend on different neuroanatomical regions, should contain tests of appropriate content (e.g., verbal vs. visuospatial) that are equivalent on item reliability and difficulty. The emergent field of rehabilitation of neuropsychological deficits will have a major problem in demonstrating that improvement in neuropsychological performance is due to the training and not to practice or recovery associated with passage of time.

Reference Notes

1. Reitan, R. M. *Manual for administration of neuropsychological test batteries for adults and children*. Unpublished manuscript, Neuropsychological Laboratory, Indiana University Medical Center, Indianapolis, 1959.
2. Diller, L., et al. *Studies in cognition and rehabilitation in hemiplegia* (Rehabilitation Monograph No. 50) Unpublished manuscript, Behavioral Sciences, Institute of Rehabilitation Medicine, New York University Medical Center, 1974.
3. Prigatano, G. P. *Neuropsychological recovery patterns for young head-injured patients with aphasia*. Paper presented at the meeting of the International Neuropsychology Society, Oxford, England, August 1977.
4. Lewinsohn, P. M., et al. *Assessment and treatment of patients with memory deficits: Initial studies*. Unpublished manuscript, University of Oregon, 1977.

References

Adams, J. On reconciling the "multidimensional" and "unitary" concepts of brain damage. *Perceptual and Motor Skills*, 1969, 29, 579-598.

- Amante, D., VanHouten, W. W., Grieve, J. H., Bader, C. A., & Margules, P. H. Neuropsychological deficit, ethnicity, and socioeconomic status. *Journal of Consulting and Clinical Psychology*, 1977, 45, 524-535.
- Barlow, D. H., & Hersen, M. Single-case experimental designs: Uses in applied clinical research. *Archives of General Psychiatry*, 1973, 29, 319-325.
- Chapman, L. F., & Wolff, H. G. The cerebral hemispheres and the highest integrative functions of man. *Archives of Neurology*, 1959, 1, 19-35.
- Chapman, L. J., & Chapman, J. P. How to test hypotheses about schizophrenia thought disorder. *Schizophrenia Bulletin*, 1975, No. 12, 42-59.
- Cohen, J., & Cohen, P. *Applied multiple regression/correlation analyses for the behavioral sciences*. Hillsdale, N.J.: Erlbaum, 1975.
- Diller, L. A model for cognitive retaining in rehabilitation. *The Clinical Psychologist*, 1976, 29, 13-15.
- Faglioni, P., Scotti, G., & Spinnler, H. Impaired recognition of written letters following unilateral brain damage. *Cortex*, 1969, 5, 120-133.
- Filskov, S. B., & Goldstein, S. G. Diagnostic validity of the Halstead-Reitan neuropsychological battery. *Journal of Consulting and Clinical Psychology*, 1974, 42, 382-388.
- Finlayson, M. A. J., Johnson, K. A., & Reitan, R. M. Relationship of level of education to neuropsychological measures in brain-damaged and non-brain-damaged adults. *Journal of Consulting and Clinical Psychology*, 1977, 45, 536-542.
- Fisher, J., & Gonda, T. A. Neurologic techniques and Rorschach test in detecting brain pathology. *A.M.A. Archives of Neurology and Psychiatry*, 1955, 74, 117-124.
- Fitzhugh, K. B., Fitzhugh, L. C., & Reitan, R. M. Psychological deficits in relation to acuteness of brain dysfunction. *Journal of Consulting Psychology*, 1961, 2, 61-66.
- Fuld, P. A., & Fisher, P. Recovery of intellectual ability after closed head injury. *Developmental Medicine and Child Neurology*, 1977, 19, 495-502.
- Gallin, D. Implications for psychiatry of left and right cerebral specialization. *Archives of General Psychiatry*, 1974, 31, 572-583.
- Geschwind, N., & Kaplan, E. A human cerebral disconnection syndrome. *Neurology*, 1962, 12, 675-685.
- Goldstein, G. The use of clinical neuropsychological methods in the localization of brain lesions. In S. J. Dimond & J. G. Beaumont (Eds.), *Hemisphere function in the human brain*. New York: Wiley, 1974.
- Goldstein, G., & Halperin, K. M. Neuropsychological differences among subtypes of schizophrenia. *Journal of Abnormal Psychology*, 1977, 86, 34-40.
- Goldstein, K. Functional disturbances in brain damage. In S. Arieti (Ed.), *American handbook of psychiatry*. New York: Basic Books, 1959.
- Gordon, N. G. Base rates and the decision making model in clinical neuropsychology. *Cortex*, 1977, 13, 3-10.

- Reitan, R. K., Baade, L. E., & Johnson, K. L. Neuropsychological test results associated with psychiatric disorders in adults. *Psychological Bulletin*, 1978, 85, 141-162.
- Robins, J. An external view of neuropsychology and its working milieu. In S. J. Segalowitz & F. J. Gruber (Eds.), *Language development and neurological theory*. New York: Academic Press, 1977.
- Rug, R. S. Antecedent probabilities, cost efficiency and differential prediction of patients with cerebral organic conditions. *Journal of Clinical Psychology*, 1971, 27, 468-471.
- Sandell, H. The use of factor scores from the Wechsler-Bellevue scale of intelligence in assessing patients with temporal lobe removal. *Cortex*, 1968, 4, 257-268.
- Schwab, P. M. *Psychological assessment of patients with brain injury*. (Report No. RD-15-P-55299/9-04) Washington, D.C.: Division of Research and Demonstration Grants, Social and Rehabilitation Services, Department of Health, Education, and Welfare, 1973.
- Schwab, P. M., Danaher, B. G., & Kikel, S. Visual imagery as a mnemonic aid for brain-injured persons. *Journal of Consulting and Clinical Psychology*, 1977, 45, 717-723.
- Strauss, A. R. *The man with a shattered world*. New York: Basic Books, 1972.
- Matarazzo, J. D., Matarazzo, R. G., Wiens, A. N., & Gallo, A. E., Jr. Retest reliability of the Halstead Impairment Index in a normal, a schizophrenic and two samples of organic patients. *Journal of Clinical Psychology*, 1976, 32, 338-349.
- Matthews, C. G., Shaw, D. J., & Kløve, H. Neuropsychological test performance in neurologic and "pseudoneurologic" subjects. *Cortex*, 1966, 2, 245-253.
- McGlone, J., & Kertesz, A. Sex differences in cerebral processing of visuospatial tasks. *Cortex*, 1973, 9, 313-320.
- Miller, W. P. Neuropsychological deficit in depression. *Psychological Bulletin*, 1975, 82, 238-269.
- Miller, B., Corkin, S., & Teuber, H. L. Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study on H. M. *Neuropsychologia*, 1968, 6, 215-234.
- Parsons, O. A. Clinical neuropsychology. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 2). New York: Academic Press, 1970.
- Parsons, O. A. Alcoholism, brain damage and altered states of unconsciousness. In M. Gross (Ed.), *Second Biannual International Symposium on Experimental Studies of Alcohol Intoxication and Withdrawal*. New York: Plenum Press, 1975.
- Parsons, O. A., & Stewart, K. D. Effects of supportive versus disinterested interviews on perceptual-motor performance in brain-damaged and neurotic patients. *Journal of Consulting Psychology*, 1966, 30, 260-266.
- Parsons, O. A., Vega, A., & Burn, J. Different psychological effects of lateralized brain damage. *Journal of Consulting and Clinical Psychology*, 1969, 33, 551-557.
- Piotrowski, Z. A. Positive and negative Rorschach organic reactions. *Rorschach Research Exchange*, 1940, 4, 147-151.
- Prigatano, G. P., & Parsons, O. A. Relationship of age and education to Halstead test performance in different patient populations. *Journal of Consulting and Clinical Psychology*, 1976, 44, 527-533.
- Reitan, R. Problems and prospects in studying the psychological correlates of brain lesions. *Cortex*, 1966, 2, 127-154.
- Reitan, R. Methodological problems in clinical neuropsychology. In R. M. Reitan & L. A. Davison (Eds.), *Clinical neuropsychology: Current status and applications*. New York: Wiley, 1974.
- Reitan, R. M., & Davison, L. A. *Clinical neuropsychology: Current status and applications*. New York: Wiley, 1974.
- Rimm, D. Cost efficiency and test prediction. *Journal of Consulting Psychology*, 1963, 27, 89-91.
- Rosenthal, R. Experimenter expectancy and the reassuring nature of the null hypothesis decision procedure. *Psychological Bulletin*, 1968, 70, 30-47.
- Satz, P. A block design rotation task: The application of multivariate decision theory analysis for the prediction of organic brain disorder. *Psychological Monographs*, 1966, 80(21, Whole No. 629).
- Satz, P., Fennell, E., & Reilly, C. The predictive validity of six neurodiagnostic tests: A decision theory analysis. *Journal of Consulting and Clinical Psychology*, 1970, 34, 375-381.
- Schachter, S. The interaction of cognitive and physiological determinants of emotional state. In P. H. Leiderman & D. Shapiro (Eds.), *Psychobiological approaches to social behavior*. Stanford, Calif.: Stanford University Press, 1964.
- Schafer, R. *Psychoanalytic interpretation in Rorschach testing*. New York: Grune & Stratton, 1954.
- Smith, A. Neuropsychological testing in neurological disorders. In W. J. Friedlander (Ed.), *Advances in neurology* (Vol. 7): *Current reviews of higher nervous system dysfunction*. New York: Raven Press, 1975.
- Spreen, O., & Benton, A. L. Comparative studies of some psychological tests for cerebral damage. *Journal of Nervous and Mental Disease*, 1965, 140, 323-333.
- Swiercinsky, D. P., & Warnock, J. K. Comparison of neuropsychological key and discriminant analysis approaches in predicting cerebral damage and localization. *Journal of Consulting and Clinical Psychology*, 1977, 45, 808-814.
- Wechsler, A. F. Presenile dementia presenting as aphasia. *Journal of Neurology, Neurosurgery, and Psychiatry*, 1977, 40, 303-305.
- Witelson, S. F. Sex and the single hemisphere: Specialization of the right hemisphere for spatial processing. *Science*, 1976, 193, 425-427.
- Zangwill, O. Intellectual status in aphasia. In P. J. Vinken & G. W. Bruyn (Eds.), *Handbook of clinical neurology*. New York: American Elsevier, 1969.

Common Methodological Problems in MMPI Research

James N. Butcher and Auke Tellegen
University of Minnesota

Research with the Minnesota Multiphasic Personality Inventory (MMPI) continues at a high rate. Unfortunately, too many articles submitted and even accepted for publication are methodologically weak. In this article we discuss some common methodological problems involving the use of the MMPI, encountered in the course of reviewing articles submitted for publication. A number of relevant issues are discussed, and some suggestions for improving research designs are made.

Research with the Minnesota Multiphasic Personality Inventory (MMPI) has continued at a high level in recent years. As a self-report inventory that includes several measures of psychopathology, the MMPI, or some derivative of it, has become widely used as a descriptive instrument or a criterion measure in a vast array of clinical and research investigations. In 1972 Buros reported that over 200 books and articles on the MMPI are published annually. Dahlstrom, Welsh, and Dahlstrom (1975) cited over 6,000 references to the MMPI; Butcher and Pancheri (1976) reported over 600 recent references in cross-national MMPI research alone. Butcher and Owen (in press) recently reviewed and classified MMPI research for the past 5 years. As shown in Table 1, they found that over one fourth of the research was focused on two areas of "popular" investigation: alcohol and drug abuse, and crime and delinquency. Approximately 37% of the studies focused on the use of the MMPI to study nonpsychiatric populations—medical patients, parents, women, ethnic groups, college students, and the aged. The MMPI seems to have been widely accepted as a criterion measure of psychopathology by researchers who want to

measure psychological problems in a variety of nonpsychiatric groups. But studies pointing to the need for revision or modification of the MMPI are scarce. This absence of revision-oriented research is impressive, since the MMPI is around 40 years old and can be expected to show some signs of aging. The only serious efforts to modify the MMPI have been directed at shortening it: About 12% of the published research in the past 5 years concerned the development or use of shorter versions of the instrument.

There are many reasons for the extensive use of the MMPI: Its administration is relatively effortless; its scoring is objective; generally straightforward objective interpretive procedures are available; and its validity as a criterion measure is comparatively well founded. But some of the factors encouraging researchers to use a self-report personality measure such as the MMPI also result in the kinds of methodological flaws that this article may help to prevent. The MMPI is so easy to administer as a part of an ongoing study that researchers (even if they have little or no research background with the instrument) feel encouraged to include it. Furthermore, since good criteria in psychopathology or personality research are scarce, the MMPI scales are sometimes too readily adopted as expedient proximate criteria. (In some respects, one might almost consider the MMPI an "attractive nuisance" in the legal sense, to be approached with appropriate caution.)

The remainder of this article will focus on

We wish to thank David Lykken for commenting on an earlier version of this article.

Requests for reprints should be sent to James N. Butcher, Department of Psychology, Elliott Hall, University of Minnesota, 75 East River Road, Minneapolis, Minnesota 55455.

some methodological problems that are often associated with research using the MMPI. We hope to provide some guidelines that could help avoid these difficulties, by providing minimal criteria. Some of these criteria emerged as a result of our experiences as reviewers of MMPI manuscripts that were submitted for publication in several psychological journals.

It will not be possible to provide an exhaustive review of all pertinent issues nor to avoid seeming at times somewhat dogmatic. Yet it is not our intention to "dictate" which areas of investigation are important or which ones should be avoided, or to impose idiosyncratic preferences for certain research methods when several defensible ones are available.

What Kind of Instrument is the MMPI?

The MMPI is often mistakenly considered to be an all-purpose personality assessment instrument that is sensitive to "normal range" personality attributes. Consequently, some researchers use the MMPI with groups for which a different instrument might be more appropriate. The standard clinical MMPI scales are measures of psychopathology, not general personality. The MMPI should not be made to do what it is not designed to do. Rather than use the MMPI inappropriately, it would be better to consider alternative or supplementary measures such as the California Psychological Inventory (CPI) or Personality Research Form, which focus on normal-range personality dimensions.

What kind of MMPI measures can or should be used? The eight clinical scales and the three validity scales plus *Mf* and *Si* are the best established and most widely used MMPI scales.¹ In view of the many problems that we have encountered in the computation of clinical scale scores, two suggestions are in order:

1. In research computations more understandable results will often be obtained if one does not use the *K* correction, but uses *K* as a separate indicator. This simply means that *K*'s validity as a "suppressor" is not assumed. Of course, appropriate multivariate analysis of one's data will permit determination whether *K* in fact does function as a suppressor in one's particular data set. The *K*

Table 1

Classification of Recent Minnesota Multiphasic Personality Inventory Research

Category	No. articles
Alcohol and drug abuse: detection and treatment	96
Short forms	66
Use with other tests or rating scales	62
Crime and delinquency	56
Diagnostic considerations: rules, profiles, and code types	50
Psychometric characteristics	49
Medical patients: physical disorders and symptoms	47
Correlation with diverse criteria	46
Women	42
Treatment variables and therapy outcome	38
Depression and suicide	38
College students and adolescents	34
Parents, couples, and families	33
Drug therapy: choice and effect	31
Race: cross-cultural, ethnic	31
Automated interpretation	26
Employment screening and job performance	24
Anxiety and stress	18
Brain damage	18
Sleep	10
Sexual deviation	9
Aging	5

Note. Adapted from Butcher and Owen (in press).

score will often be found to be an additional measure of good adjustment rather than a measure of invalid variance. If non-*K*-corrected *T* scores are used, use the *correct* conversion tables. Occasionally manuscripts are received containing figures with non-*K*-corrected profiles drawn according to incorrect norms. Do not use the raw score indications printed on the standard profile sheet unless your scores are actually *K* corrected! If you are plotting non-*K*-corrected scores for Scales *Hs*, *Pd*, *Pt*, *Sc*, and *Ma* and plan to display these on the standard profile sheet, it is necessary to obtain the correct *T*-score elevation from a conversion table (e.g., in Dahlstrom, Welsh, & Dahlstrom, 1972) and draw the appropriate elevation from the *T*-score indica-

¹ Researchers interested in obtaining access to the item response data of the original Minnesota standardization sample should contact W. Grant Dahlstrom, Department of Psychology, University of North Carolina, Chapel Hill, North Carolina 27514.

tions on the side of the profile sheet. In spite of the above general recommendation about using non-*K*-corrected scores, it is sometimes useful for comparisons with other studies to do a parallel set of analyses using *K*-corrected scores.

2. Use raw scores in research computations rather than *T* scores unless the use of *T* scores is specifically indicated, for example, in the analysis of code types or when combining data from both sexes.

The eight clinical scales were developed primarily by means of empirical keying methods, using clinical diagnoses as criteria. Because of the primary reliance on external criteria rather than on internal structure, these scales are for the most part quite heterogeneous. Consequently, although these scales, particularly when used in combination, provide useful information as to the diagnostic status of a person (Dahlstrom et al., 1972; Overall, Note 1), they do not in any simple way reflect "what the patient is saying," that is, how the patient is *describing herself/himself* through the medium of the MMPI items. This, too, is important information. Nowadays we recognize more clearly than in the heyday of empirical keying and "dust bowl empiricism" that important and clinically relevant *direct* relations exist between the manifest content of a person's self-description and clinical symptomatology and behavior in general (e.g., Hase & Goldberg, 1967; Jackson, 1971; Koss & Butcher, 1973; Payne & Wiggins, 1972). In addition, in clinical use the client/patient's reply to MMPI questions, impersonal as the process may seem, is still a communication that he/she might expect the psychologist to receive. We therefore need to take steps to make this possible. Some organized way of "coding" the respondent's MMPI "message" is clearly called for. Two approaches are now open:

1. use of content homogeneous scales, including factor or cluster scales (Block, 1965; Eichman, 1962; Overall, Hunter, & Butcher, 1973; Stein, 1968; Welsh, 1956; Wiggins, 1969) and

2. use of so-called "critical items" (Grayson, 1951; Koss & Butcher, 1973; Koss, Butcher, & Hoffmann, 1976).

Both approaches can serve a very useful role complementary to that of the clinical scales.

Scale Proliferation

The development of additional MMPI scales has been a rather popular pastime for psychologists. There are presently more MMPI scales than there are items on the inventory! Unfortunately, much of the MMPI scale development does not derive from a sound conceptual framework. Many scales have been constructed by contrasting different "samples of convenience" (often of heterogeneous makeup or with important characteristics unknown). Often these scales are not cross-validated, and more often than not their psychometric properties and interrelations with other scales are not reported. Many new scales, since they are, after all, derived from the same item pool, prove to be largely redundant alternative versions of existing scales, although sometimes of poorer quality.

A researcher who is tempted to add to this plethora of scales would find his or her contribution to the MMPI literature better received and used if:

1. The investigator has very carefully considered the question of whether the MMPI item pool adequately covers the domain of content of the construct to be measured. If the answer is not affirmative, then the MMPI will obviously not serve the investigator's purpose very well, and the new scale will reflect more the characteristics of the MMPI item pool than the domain of the intended construct. We believe that this is in fact the case with some new MMPI scales.

2. The investigator can show the scale to be conceptually interesting.

3. The scale is developed and cross-validated on reasonably large, well-defined samples (whether it is an empirically keyed scale or one derived by internal consistency methods, e.g., factor analysis).

4. The internal structure of the scale is reported (internal consistency, possibly factor structure), and the relation to other MMPI scales is reported. These relations along with validity data would ideally reveal that the scale either is a superior alternative to an

existing MMPI scale or is a measure of a disposition or state not measured by existing scales.

5. In addition to publishing statistics on scale relationships, the author reports substantial external validation data. It is not sufficient to know that a scale correlates appreciably with a criterion, since, ideally, both convergent and discriminant validity information should be reported (Campbell & Fiske, 1959) for the new scale. It is also important to know how well a scale predicts behavior for the individual in various contexts. Thus some data reporting prediction success and failure in specific settings (e.g., in the form of hit-miss tables) are desirable.

New MMPI scales, like many of the old ones, should not be assumed to measure the characteristics suggested by its name or by its author. The researcher should be aware of the track record of a particular scale before he or she gambles on it. Block (1965) provided an interesting illustrative example:

However, ease of analysis should not mean casualness in regard to the scales chosen for study. On occasion, scales have been employed in circumstances which can only mean ignorance or naivete. These are harsh words but consider the following: several correlational studies in the response set domain have employed both the Prejudice (Pr) and Tolerance (To) scales of the MMPI. Both of these scales are due to Gough (1951; 1952). The Pr scale was developed and validated as an MMPI measure to correlate with the California Ethnocentrism-Fascism Scale (Adorno et al., 1950). Later, Gough decided to revise the Pr scale slightly and for entirely appropriate conceptual reasons took the occasion to relabel the scale as a measure of tolerance, reversing the direction of scoring of the items. In the To scale, 29 of the 30 items overlap with the Pr scale but scored oppositely.

Since the sample of scales selected for study can determine the shape of the results obtained, simple considerations of availability must be bolstered by actual knowledge of candidate scales. To use both the Pr and the To scales in the same study and draw faulty conclusions from their opposite but identical factor loadings is informative only about the investigators. (p. 118)

The MMPI wears other clothes as well. Many scales in the personality research domain have been generated from the MMPI item pool or from constructs so close to those covered by the MMPI that the items are virtually the same. Often the new researcher

is unaware of the original relationships or similarities and then conducts an "empirical" study rediscovering the affinity of the tests in question. A few instruments that are wholly or in part derived from the MMPI or are closely related are the CPI (over 200 items are common), the Taylor Manifest Anxiety Scale, and Lanyon's Psychological Screening Inventory.

MMPI Short Forms

A large number of recent studies have used one of the several MMPI short forms. It should be noted that the only MMPI short form recognized by the test authors and the publishers consists of a reduced number of items (around 400) that include *all* of the items required to score the 3 validity and 10 standard scales. The items excluded are those at the end of the booklet that are not scored on the basic scales. Some short forms have been the object of a large amount of recent MMPI work and have been developed by varying methods and for different purposes (the Mini-Mult, Midi-Mult, the Faschingbauer Abbreviated MMPI, MMPI-168, etc.)

Several recent studies have pointed out the limitations of MMPI short forms (Fillenbaum & Pfeiffer, 1976; Hedlund, Won Cho, & Powell, 1975; Hoffmann & Butcher, 1975). Although MMPI short forms may correlate significantly with the full MMPI, the resulting code-type congruence (hit rate) between the two forms is quite low (from 33% to 49%), too low to result in very similar individual clinical decisions. Although these studies indicate the need for caution in not accepting short forms at face value as adequate substitutions for full-length MMPI measures, many investigators continue to consider an MMPI short form as a near-equivalent set of scales. It is conceivable that some short forms could serve in a much more limited way as measures of global psychopathology.

How to Find Relationships: Problems in Discovering and Appraising Relations Between the MMPI and Other Variables

All MMPI research can be viewed as essentially correlational research, whether the

MMPI scales are the "dependent" or "independent" variables and whether we compute correlations explicitly or, for example, do *t* tests, comparing one group with another on certain MMPI measures (e.g., patients vs. normals), or determine hit rates obtained with certain MMPI-based diagnostic rules. Sometimes the *strength* of the relation investigated is immediately given by our statistics (e.g., correlation coefficient, hit rate), sometimes it is not (*t* test, *F* ratio). It is known but easily forgotten that some measure of strength of relation is *always* necessary if we are to determine the importance of our findings. Demonstrating statistical significance is necessary, but it is not enough. Not only should the magnitude of, for example, a correlation coefficient be considered but also its confidence interval. Yet one suspects that an inverse relation exists between the sample sizes used in published studies and the strength of the relations that are claimed to exist between certain variables. The reason, of course, is that a good way of "discovering" dramatically strong relationships in a set of data is to collect several measures, preferably those that can be expected to show *some* moderate interrelations, using a small sample. Chance alone will move a few of these expectable relations up into the "strikingly high" range (the "trade-off" being that one can expect to "lose" some of the other relationships by fluctuations in the opposite direction).

In short, sample sizes should be large enough to permit conclusions about the *strength* of the relationships under consideration that go beyond the inference that the null hypothesis is false. The rejection of this null hypothesis is a trivial contribution, particularly if the observed relationship is in the (more or less) expected direction, because, as has been pointed out, we practically *know* the null hypothesis to be false—almost everything is somewhat related to everything (Lykken, 1968; Meehl, 1967). *Example:* Suppose one collects data on a (small) sample of 19 subjects and "discovers" a somewhat unexpected correlation of .50 between Scale 7 and a psychophysiological measure of "arousal," skin conductance level. The corresponding *z* value of .50 is .55 with a standard error of .25.

The .05 (two-tailed) confidence interval for this correlation, then, in *z* values, is $.55 \pm .49$. In correlational values the corresponding range is between .06 and .78. The finding of a possibly very weak correlation, one that is only significantly greater than .06, we submit, is hardly a contribution (particularly not in this case, since the direction of the relation was more or less anticipated), in spite of its statistical significance and observed magnitude. Large sample sizes are essential to achieve less trivial results. Admittedly, if an observed relation is definitely not in the expected direction, and is definitely "significant" (even though the estimate of its strength still has a large margin of error), and involves a truly interesting relationship, then we might speak of a finding that deserves publication at this point to stimulate further research in what now appears to be a promising area. Nevertheless, it would still be better to follow up one's own findings with a replication permitting a more precise estimate and demonstrating a relationship of nontrivial strength. (After all, the more interesting our findings are to us, the less expected they must have been, and therefore the less reason we apparently have to assume their replicability without further test!)

One fairly common practice is to calculate multiple correlations, often on a small sample, and to announce their statistical significance without reminding the reader that the magnitude of the obtained multiple correlations may be substantially inflated. Without cross-validation or an appropriate "shrinkage" estimate, this procedure can be grossly misleading. (See Schmitt, Coyle, & Rauschenberger, 1977, for a recent discussion of this topic.)

Almost *all* MMPI research is multivariate (whether or not multivariate statistics are actually used). This, in combination with the frequent availability of "samples of convenience" (consisting, e.g., of individuals who have completed the MMPI as part of a screening procedure in some clinical or counseling setting), inevitably invites "look-see" or "shotgun" studies, guided only by the investigator's vague hope of finding something somewhere among the many possible relations that might be subjected to scrutiny. Particularly if the sample is small, chance findings

will often occur, and if one's sample of convenience is, in addition, poorly defined, then even nonchance findings may well prove not replicable in the absence of clear understanding of to what population the results are to be generalized. Unfortunately, a rather large number of manuscripts describing "one shot" studies using small and haphazardly collected samples continue to be submitted for publication—and to appear in print.

Another undesirable practice is to boost sample sizes by lumping together subjects differing on some potentially important characteristics such as sex, age, race, and socioeconomic status (Carlson, 1971). Sometimes these variables may in fact prove not to be important for the relations of interest to the investigator. But this cannot be assumed. At the very least sex ought to be included among one's variables. However, this would still not allow for the possibility that relationships between the MMPI and other variables, for example, response to treatment or diagnostic status, might differ for the two sexes, for different age groups, and so on. If sample sizes permit, the simplest first step would be to subdivide one's sample into relatively homogeneous subgroups in respect to major differentiators (sex, age, and so forth) and to evaluate the relationships in the different subgroups for possible differences.

Because of the MMPI "code type" tradition, it is not uncommon to analyze relationships by presenting MMPI correlates in the form of relevant base rates or means associated with each of several distinct code types. (The actuarial atlases of Marks, Seeman, & Haller, 1974, and Gilberstadt & Duker, 1965, are well-known examples.) One distinct potential advantage of this approach is that certain "configural" relations, reflecting interactions among MMPI scales in relation to criterion variables, will be used. For example, if correlates of Scale 7 differ depending on the subject's scores on additional scales, say 2, 4, or 8, then the use of different code types such as 2-7, 4-7, 2-7-8, and so on, permits one to capitalize on these interactions.

But reliance on code types has disadvantages. One is that some information is inevitably lost when a profile is assigned to a code type. Another disadvantage is the small

number of classifiable subjects for many code types, even in rather extensive studies. These classification problems are often unnecessary, since many important relations between MMPI and external criteria are often not configural but essentially linear. That is, important relations may be adequately described by the linear relations existing in the *total* sample between one or more MMPI variables and other measures. One example is Goldberg's (1965) analysis of methods of distinguishing psychotic from neurotic patients on the basis of the MMPI. Goldberg showed that a simple linear combination of several clinical scales did just as well as the highly configural Meehl-Dahlstrom rules. Although the Goldberg analysis is one of the few well-documented examples, we have little doubt that (e.g., in the material from which Marks et al. 1974, and Gilberstadt & Duker, 1965, atlases have been derived) a number of linear relations between single or several MMPI scales and important nontest data are hidden, other than the psychotic versus neurotic distinction. One example is an apparent straightforward (linear) relationship between scores on Scale 4 and alcoholism in the Gilberstadt and Duker (1965) sample (pointed out by them). This example illustrates another disadvantage of exclusively relying on the "code-type approach": The method may *exploit* but does not *reveal the nature of* the relations, linear or nonlinear, that may exist in one's data.

Goldberg's monograph is one example of how such relations can be pursued systematically, albeit in this case with negative results with respect to nonlinear relations. However, we believe that even nonlinear relations may emerge from systematic analyses provided the search is guided by a certain degree of theorizing permitting a focus on a smaller set of possible configurations, thus reducing the probability of drowning a few actual configural relations in a sea of unreplicable chance relationships. The reason is that without some theoretical constraints, the number of configural patterns that would have to be considered in an exhaustive search easily becomes extremely and unmanageably large, thus requiring unattainably large sample sizes to minimize the occurrences of chance patterns.

We find, then, that code and profile types provide convenient ways of assimilating the information contained in a set of MMPI scores. But in spite of the continuing advocacy (and plausibility) of typological classifications, acceptance of such classifications should be provisional rather than uncritical. We still do not know a great deal about how much and when type of membership would add to the information already contained in a (more parsimonious and psychometrically conservative) dimensional predictor, for example, an appropriate linear regression formula. Research on this question seems quite feasible and could give interesting results.

Assessing Profile Change

The measurement of change is methodologically complex, and the researcher engaged in such measurement should become aware of the difficulties involved (Cronbach & Furby, 1970; Fiske & Rice, 1955; Mauger, 1972; see especially Dahlstrom et al., 1972, chap. 7). We cannot review all of the relevant issues in the limited space of this article. However, it is important to keep in mind that a simple difference in MMPI mean scale scores on retest may not indicate that the individuals involved have necessarily changed as a result of some interposed treatment. Familiarity with and critical understanding of such concepts as "residual gain" are essential (e.g., Cronbach & Furby, 1970). To balance somewhat the note of pessimism in Cronbach and Furby's article, it should be pointed out that experimental designs containing comparison groups can be effective tools for understanding change. One substantive issue that the investigator should keep in mind is that MMPI measures appear to tap a mixture of changeable and stable characteristics. Some of the scales (e.g., *D*) appear to be reflective of "state" as well as "trait" characteristics and are sensitive to mood changes. Other scales focus on "biographical" factors (*Si* or *Pd*) and seem relatively more stable. What is the meaning of a particular MMPI scale change? If not a statistical artifact, does it reflect a change in affective state, a permanent change, a change in self-presentation?

In addition, a change on a particular scale

is not necessarily best described in terms of the name of the scale in question. For example, a significant change on the *Pd* scale may be due primarily to changes on a subset of *Pd* items reflecting a negative mood state; content-homogeneous MMPI scales are certainly less potentially misleading in the interpretation of change than the more heterogeneous clinical scales. These problems should be addressed in any thoughtful discussion of MMPI data pertaining to change. As a general point, it is good to keep in mind that some scales may be useful for *reflecting* change, say, resulting from intervention, but other scales may be more useful for *predicting* response to treatment (without necessarily reflecting the effects of treatment).

On the whole it may be true that the original method of developing the MMPI favored the selection and construction of relatively stable items and scales, respectively. None of the original work was directly focused on measuring change. Subsequent efforts at using MMPI items to detect changeability or to develop scales that would give clues to potential for change have not met with much success (Mauger, 1972; Pepper, 1964). The "stability" of the MMPI is shown by the fairly consistent finding that about 87% of the MMPI items are answered in the same direction on retest (Butcher & Gur, 1974; Goldberg & Jones, 1969; Schofield, 1948; Ullmann & Wiggins, 1962). Consequently, the number of items that can be expected to vary on retest is relatively small.

Reporting Group Data

What do we know from a group mean profile that summarizes a set of MMPI scores? Can we assume that if the group average is a 278" profile type that the established personality correlates for the 278" profile necessarily fit the group as a whole or some subset of individuals in the group? No, sometimes we fall into the trap of drawing such a conclusion. It may actually be the case that *no* individual in the group has the code type corresponding to the group mean score. Researchers who are not content to present merely the means and standard deviations of individual scales could report the percentages

of the different code or profile types occurring in their sample.

In summary, the MMPI continues to be a widely used research instrument as well as a clinical assessment inventory. Reasons for its wide application include its easy administration, its objective scoring, the volume of data demonstrating its validity, and the automation of its interpretation. The MMPI is such an easily used research instrument that it is sometimes misapplied or the data obtained with it are incorrectly analyzed by researchers who are not familiar with some of its limitations or peculiarities. The MMPI can play and has played a useful role in significant research. The main prerequisites are the investigator's long-term and genuine concern with an important problem and her/his ability to draw on the store of substantial methodological knowledge and wisdom, accumulated in the field of personality assessment in general (e.g., Wiggins, 1973) and in the area of MMPI research in particular (e.g., Dahlstrom et al., 1972, 1975).

Reference Note

1. Overall, J. E. *Implementation of an actuarial diagnostic program in a clinical setting*. Paper presented at the 11th annual symposium on recent developments in the use of the MMPI. Minneapolis, April 1976.

References

- Block, J. *The challenge of response sets: Unfounding meaning, acquiescence and social desirability in the MMPI*. New York: Appleton-Century-Crofts, 1965.
- Buros, O. *Seventh mental measurements yearbook*. Highland Park, N.J.: Gryphon Press, 1972.
- Butcher, J. N., & Gur, R. Hebrew translation of the MMPI: An assessment of translation adequacy and preliminary validation. *Journal of Cross-cultural Psychology*, 1974, 5, 220-227.
- Butcher, J. N., & Owen, P. L. Objective personality inventories. In B. B. Wolman (Ed.), *Clinical diagnosis of mental disorders: A handbook*. New York: Plenum Press, in press.
- Butcher, J. N., & Pancheri, P. *Handbook of cross-national MMPI research*. Minneapolis: University of Minnesota Press, 1976.
- Carlson, R. Where is the person in personality research? *Psychological Bulletin*, 1971, 75, 203-219.
- Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Cronbach, L. J., & Furby, L. How we should measure "change"—Or should we? *Psychological Bulletin*, 1970, 74, 68-80.
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook* (Vol. 1). Minneapolis: University of Minnesota Press, 1972.
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook* (Vol. 2). Minneapolis: University of Minnesota Press, 1975.
- Eichman, W. J. Factored scales for the MMPI. *Journal of Clinical Psychology*, 1962, 15, 363-395.
- Fillenbaum, G. G., & Pfeiffer, E. The Mini-Mult: A cautionary note. *Journal of Consulting and Clinical Psychology*, 1976, 44, 698-703.
- Fiske, D. W., & Rice, L. Intra-individual response variability. *Psychological Bulletin*, 1955, 52, 217-250.
- Gilberstadt, H., & Duker, J. *A handbook for clinical and actuarial MMPI interpretation*. Philadelphia, Pa.: Saunders, 1965.
- Goldberg, L. R. Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs*, 1965, 79(9, Whole No. 602).
- Goldberg, L. R., & Jones, R. R. The reliability of reliability: The generality and correlates of intra-individual consistency in responses to structured personality inventories. *Oregon Research Institute Monograph*, 1969, 9, No. 2.
- Grayson, H. M. *A psychological admissions testing program and manual*. Los Angeles: Veterans Administration Center, Neuropsychiatric Hospital, 1951.
- Hase, H. D., & Goldberg, L. R. Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin*, 1967, 67, 231-248.
- Hedland, J. L., Won Cho, D., & Powell, B. J. Use of MMPI short forms with psychiatric patients. *Journal of Consulting and Clinical Psychology*, 1975, 43, 924.
- Hoffmann, N. G., & Butcher, J. N. Clinical limitations of three MMPI short forms. *Journal of Consulting and Clinical Psychology*, 1975, 43, 32-39.
- Jackson, D. N. The dynamics of structured personality tests: 1971. *Psychological Review*, 1971, 78, 239-248.
- Koss, M. P., & Butcher, J. N. A comparison of psychiatric patients' self-report with other sources of clinical information. *Journal of Research in Personality*, 1973, 7, 225-236.
- Koss, M. P., Butcher, J. N., & Hoffmann, N. G. The MMPI critical items: How well do they work? *Journal of Consulting and Clinical Psychology*, 1976, 44, 921-928.
- Lykken, D. T. Statistical significance in psychological research. *Psychological Bulletin*, 1968, 70, 151-159.
- Marks, P. A., Seeman, W., & Haller, D. *The actuarial use of the MMPI with adolescents and adults*. Baltimore, Maryland: Williams & Wilkins, 1974.
- Mauger, P. A. The test-retest reliability of persons: An empirical investigation utilizing the Minnesota Multiphasic Personality Inventory and the Per-

- sonality Research Form. (Doctoral dissertation, University of Minnesota, 1972.) *Dissertation Abstracts International* 1972, 33, 2806B. (University Microfilms No. 72-32, 328)
- Meehl, P. E. Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 1967, 34, 103-115.
- Overall, J. E., Hunter, S., & Butcher, J. N. Factor structure of the MMPI-168 in a psychiatric population. *Journal of Consulting and Clinical Psychology*, 1973, 41, 284-286.
- Payne, F. D., & Wiggins, J. S. MMPI profile type and the self-report of psychiatric patients. *Journal of Abnormal Psychology*, 1972, 79, 1-8.
- Pepper, L. J. *The MMPI: Initial test predictors of retest changes*. Unpublished doctoral dissertation, University of North Carolina, 1964.
- Schmitt, N., Coyle, B. W., & Rauschenberger, J. A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlations. *Psychological Bulletin*, 1977, 84, 751-758.
- Schofield, W. *MMPI response changes with certain therapies*. Unpublished doctoral dissertation, University of Minnesota, 1948.
- Stein, K. B. The TSC scales: The outcome of a cluster analysis of the 550 MMPI items. In P. McReynolds (Ed.), *Advances in psychological assessment*. Palo Alto, Calif.: Science & Behavior Books, 1968.
- Ullmann, L. P., & Wiggins, J. S. Endorsement frequency and the number of differentiating MMPI items to be expected by chance. *Newsletter of Research in Psychology*, 1962, 4, 29-35.
- Welsh, G. S. Factor dimensions A and R. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: University of Minnesota Press, 1956.
- Wiggins, J. Content dimensions in the MMPI. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications*. New York: McGraw-Hill, 1969.
- Wiggins, J. S. *Personality and prediction: Principles of personality assessment*. Reading, Mass.: Addison-Wesley, 1973.

Received September 26, 1977 ■

Methodological and Interpretive Problems of Single-Case Experimental Designs

Alan E. Kazdin
Pennsylvania State University

Single-case experimental designs have been used with increasing frequency in clinical research. These designs are uniquely suited to evaluating treatment effects with individual clients. Although treatment is evaluated by comparing baseline and treatment phases, the manner in which this is accomplished varies as a function of the specific design. Typically, the comparison is replicated over time (ABAB design) or across different behaviors (multiple-baseline design). Several methodological problems frequently arise in single-case designs, such as deciding when to alter phases or conditions in the experiment, ensuring that the intervention is implemented, comparing alternative treatments unconfounded by sequence effects, and ensuring that data are collected reliably. Many interpretive problems of single-case designs stem from the criteria used to evaluate treatment. Whether treatment produced a reliable effect usually is determined by visual inspection of the data. In addition, single-case research has been concerned with the clinical importance of treatment effects. The ambiguity of these criteria, relative to statistical tests used in group designs, presents unique problems for evaluating treatment. The present article considers methodological and interpretive problems that frequently arise in single-case experiments.

Single-case experimental designs are used increasingly in clinical research. Although these designs can evaluate interventions with groups of subjects, their unique contribution is that they can experimentally evaluate interventions for the individual client. Hence, the designs offer a distinct methodological advance over the traditional uncontrolled case study, which provides only suggestive information about the effects of treatment on a client's behavior.

Several characteristics and problems of single-case experimental designs are still un-

familiar to many investigators who might profit from their use. The designs and their execution are not inherently complex. However, they do differ in important ways from traditional group research with which most investigators are familiar. The present article describes several methodological problems that frequently arise in single-case experiments. These encompass planning the research, making decisions during the investigation, and evaluating the results. To understand these methodological problems, the basic rationale of single-case experiments and the manner in which intervention effects are demonstrated need to be discussed briefly.

This manuscript was completed while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, California. Fellowship support was obtained from the Foundations Fund for Research in Psychiatry, National Institute of Mental Health Grant 1-T32-MH14581, and the Spencer Foundation, which the author gratefully acknowledges.

Requests for reprints should be sent to Alan E. Kazdin, Department of Psychology, Pennsylvania State University, University Park, Pennsylvania 16802.

Overview of Single-Case Experimental Designs

Basic Rationale

The rationale of single-case designs is similar to traditional group research. In traditional experimentation, the effect of an intervention is assessed by comparing performance

under the influence of different levels of a given variable. Typically, different groups of subjects or the same group are exposed to different experimental conditions. The essential feature of traditional research, as well as single-case research, is a comparison of performance under different conditions. The manner in which the comparison is made in single-case experiments departs from traditional designs.

Single-case experimental designs usually begin by observing a client's behavior before treatment for a period of at least several days. This period, referred to as a baseline phase, serves a twofold purpose. First, the data collected during baseline describe the existing level of performance and hence provide information about the extent of the client's problem. Second, the data serve as a basis for predicting the level of performance for the immediate future if treatment is not provided. Even though the descriptive function of baseline data is important for characterizing the extent of the problem, from the standpoint of design, the predictive function is central.

To evaluate treatment, it is important to have a clear idea of what behavior would be like in the immediate future with no intervention. A projection of baseline performance into the future is the implicit criterion against which treatment is evaluated. Thus, a stable level of performance during baseline is important before beginning treatment. A stable rate of performance is characterized by the absence of trend (slope) in the data and only slight or moderate variability in performance. Once a stable rate is obtained, treatment can be implemented. Assessment of behavior is continued while treatment is in effect to determine whether performance departs from baseline levels. If treatment is effective, the actual level of behavior should deviate from the projected level of behavior from baseline performance. Data during the treatment phase serve another function. These data also provide a new level of performance and can be used to predict how the client will behave in the immediate future if treatment is continued. After performance stabilizes, the treatment can be withdrawn to reassess whether performance under these conditions deviates

from the predicted level. Also, by withdrawing treatment, the investigator tests directly whether the original baseline level of performance would have continued at a given level.

Essentially, data in separate phases of single-case designs provide information about present performance, predict the probable level of future performance, and test the extent to which predictions of performance from previous phases were accurate. By repeatedly altering experimental conditions in the design, there are several different opportunities to compare phases and to test whether performance is altered by the intervention. If performance changes in response to alterations of the experimental conditions, the change can be more parsimoniously accounted for by the experimental conditions than by extraneous events.

Illustrative Designs

The manner in which treatment effects are demonstrated varies as a function of different designs. There are several designs, each of which includes many variations (see Baer, Wolf, & Risley, 1968; Hartmann & Hall, 1976; Hersen & Barlow, 1976; Kazdin, 1978b; Leitenberg, 1973; Ulman & Sulzer-Azaroff, 1975). Two designs warrant brief mention because they are the most commonly used in single-case experimentation and because they serve as a useful point of departure for describing methodological problems for single-case experiments. These designs are the ABAB, or reversal, design and the multiple-baseline design.

The ABAB design evaluates an intervention by alternating the baseline condition (A phase) when no treatment is in effect with the intervention (B phase). Baseline data are collected until the client's performance stabilizes. At this point, the intervention is implemented. When performance again stabilizes or shows a trend that clearly departs from the projected performance of baseline, the intervention usually is withdrawn (second A phase). Finally, as performance shifts to a new level, the intervention is reinstated (second B phase). Typically, behavior again changes in response to the intervention. Systematic changes in behavior associated with

variation of the experimental conditions, particularly when replicated at different points in the design, strongly suggest that the intervention accounts for the results.

The multiple-baseline design, used less frequently than the ABAB design, demonstrates the effect of the intervention without withdrawing treatment. In this design, the effect of the intervention is demonstrated by showing that behavior change accompanies introduction of the intervention at different points in time. Although there are different versions of the design, a commonly used version is the multiple-baseline design across behaviors. In this version, baseline data are gathered across two or more different responses of a single subject (or group of subjects). After each behavior shows a stable rate, the intervention is applied to only one of the behaviors. Baseline conditions remain in effect for the other behavior(s). Typically, the behavior to which the intervention was applied changes, whereas the other behaviors remain at baseline levels. When all responses show a stable rate, the intervention also is applied to the second behavior; remaining behaviors continue under baseline conditions. This procedure is continued so that the intervention is introduced to one response at a time. A causal relation between the intervention and behavior is demonstrated if each response changes when and only when the intervention is introduced. If a change in behavior at the point that the intervention is introduced is replicated across several behaviors, this provides a convincing demonstration that the intervention was responsible for behavior change.

There are other designs that demonstrate intervention effects in ways that vary in detail from the ABAB and multiple-baseline designs described briefly above. In each of these designs, the fundamental comparison to evaluate treatment is the performance of baseline and intervention phases. The comparison usually is based on replications within a single subject so that repeated opportunities are available to assess whether the intervention accounted for change. It is important to understand the basic rationale for evaluating treatment that transcends the diverse designs, because ambiguities in results of single-case

experiments frequently derive from not meeting the conditions required by the designs.

Common Methodological Problems

Several methodological problems are common in investigations using single-case experimental designs. Many problems are independent of the specific designs that are used. Hence, it is not important to fully detail the specific designs available in single-case experimentation as long as the rationale is illustrated with samples of the most frequently used designs. Areas in which problems frequently arise are altering phases in the designs, checking whether the intervention was manipulated correctly, comparing alternative interventions, and gathering reliable data on client behavior.

Altering Phases or Conditions During the Experiment

Traditionally, research designs are preplanned so that most of the details are arranged before subjects are studied. In single-case experimental designs, a number of important decisions can be made only as the data are collected. Decisions such as how long baseline data should be collected and when to present or withdraw experimental conditions are made during the investigation itself. The experimenter needs to decide when to alter phases in the design in such a way as to maximize the clarity with which inferences can be drawn about the intervention. Each of the single-case designs usually begins with a baseline phase. Treatment is evaluated ultimately by comparing performance across baseline and intervention phases, as described earlier. For these comparisons to be made easily, one has to be sure that the changes from one phase to another are likely to be due to the intervention rather than to a continuation of an existing trend. A fundamental design issue is when to change phases to maximize the clarity of data interpretation.

There are no clear rules for altering phases. Understandably, investigators differ when conditions are altered. Yet, the point at which conditions are changed is a very important issue, because subsequent evaluation of inter-

vention effects depends completely on how clear the behavior changes are across phases. The usual rule of thumb is to alter phases in which the data are stable. Stability refers to the absence of trend and relatively small variability in a given level of performance. Trends and extensive variability during any of the phases, particularly baseline, can interfere with evaluating treatment.

Trends in the data. Ideally, baseline data should show no trend or slope prior to implementing the intervention. One of two different trends may be apparent. First, behavior may be changing in the direction opposite from that which is to be achieved during the intervention. For example, a psychotic patient may show a reduction in rational statements during baseline. Because the intervention will attempt to alter behavior in the opposite direction (increase rational statements), this initial trend in baseline is not likely to interfere with subsequent conclusions about the intervention. As a general rule, when the intervention is designed to change behavior in a direction opposite from the trend in baseline, the trend is not problematic. The rule can be extended to all changes in conditions so that trends, opposite from what the anticipated data will show in the next phase, should not interfere with drawing inferences about treatment.

In contrast, the baseline trend may be in the same direction that the intervention is likely to produce. Essentially, baseline may show improvements in behavior, and one might question the need to intervene at all. However, even if improvements are made, as might be the case with an autistic child's severe self-injurious behavior during a baseline period when attention for such behavior is inadvertently withheld, the changes may be so slow that some intervention is needed. For example, a child engaging in frequent head banging might gradually decrease the response, but the reduction could be so gradual that serious self-injury might be inflicted unless the behavior is quickly eliminated. Despite the desirability of intervening in many situations in which baseline trends move in the direction of therapeutic change, evaluating the effect of the intervention in these situations is extremely difficult. The in-

tervention has to produce very marked change to draw unambiguous conclusions. Because of this difficulty in evaluating interventions with systematic trends in the direction of therapeutic change in baseline, the usual recommendation is to wait for baseline to stabilize so that there is no trend before intervening (Baer et al., 1968). This cannot be done in many clinical situations in which treatment is needed quickly.

When baseline shows a trend toward improvement, two major alternative strategies are available. First, an ABAB design can be used in which a procedure for changing behavior in the opposite direction can be alternated with the intervention. For example, baseline can consist of reinforcing *decreases* in the desirable behavior, whereas interventions can reinforce *increases* in the same behavior. This alternative can dramatically demonstrate that treatment accounted for change (e.g., Ayllon & Haughton, 1964). The design is experimentally sound but clinically untenable because it includes specific provisions for making the client's behavior worse.

A second solution is to select designs in which trends in the data are unlikely to affect each of the baselines observed (multiple-baseline design) or in which trends are not relevant to evaluate different treatments (simultaneous treatment design). A third solution is to use statistical techniques that can evaluate the effect of the intervention relative to the baseline trend. Specific statistical techniques such as time-series analysis can take into account baseline trends and assess whether the intervention has made a reliable change over and above what would be expected from continuation of the trend (see Glass, Willson, & Gottman, 1974; Jones, Vaught, & Weinrott, 1977; Kazdin, 1976).

Intrasubject variability. In addition to trends, excessive variability in the data can interfere with drawing conclusions about treatment. As a general rule, the greater the variability in the data, the greater the difficulty in demonstrating behavior change. Excessive variability is a relative matter that depends on the initial level of behavior during baseline, the magnitude of change achieved during the intervention, the degree of change desired, and other factors. In the

extreme case, baseline performance of the target behavior may vary daily from 0 to 100%, that is, to each extreme of an assessment continuum. Such baseline data probably cannot be used to predict a specific level of future performance and hence would not provide an adequate base against which the data during an intervention phase could be compared.

To facilitate treatment evaluation, investigators sometimes reduce the *appearance* of variability in graphical presentation of the data by averaging data points across consecutive days or sessions. The procedure consists merely of combining days of data and averaging across the number of combined data points. By representing 2 or more days with a single averaged data point, fluctuations are reduced substantially and the data appear more stable. Although such a procedure distorts the day-to-day pattern of performance, the data in different phases can be readily compared.

Whenever possible, it is better to identify and control sources producing variability than merely to average the data. The variability may result from sources that are important to identify before implementing treatment. For example, unreliability in scoring behavior may contribute to excessive variability. The client may perform the target response relatively consistently. Yet observers may change criteria in scoring the behavior, which may add considerably to the variability. Such variability should be controlled by ensuring that observations are scored consistently across observers and over time. Other sources may account for variability, such as changes in cues in the environment. For example, the presence of one individual rather than another (e.g., mother rather than father) may control the rate of behavior. If variability is excessive, an attempt can be made to assess the stimulus conditions that mediate differences in response rates. Identification of some of the factors that contribute to variability also may provide cues about the factors controlling behavior and provide practical suggestions for treatment.

For whatever reason, behavior may simply be quite variable. Indeed, the goal of the program may be to alter the variability of the

client's performance rather than the mean rate. Usually, large data fluctuations, especially during baseline, make evaluation of the intervention difficult. Such fluctuations are a signal to the investigator to either look for the controlling variables or to redefine the behavior or conditions of assessment to reduce variability. Continuation of the program in the face of excessive variability eliminates the possibility of drawing inferences about treatment.

Duration of the phases. How long phases will be in a single-case experimental design usually is not specified in advance of the investigation. The reason for this is that the investigator needs to examine the data and to determine whether the information is sufficiently clear to make predictions about performance. The presence of trends or excessive variability indicates that the information may not be sufficiently clear. Hence, it is useful to wait for a more stable rate of performance to emerge. A common methodological problem is altering phases before a clear pattern emerges. For example, most of the data may indicate a clear pattern for baseline. Yet, after a few days of relatively stable baseline performance, one or two data points may be higher or lower than the previous data. The question that immediately arises is whether a trend is emerging in baseline or whether the data points are merely a part of the normal variability. It is wise to continue the condition without shifting phases just to be sure. If 1 or 2 more days of data reveal that there is no trend, the intervention can be implemented as planned. The increase of data provides an increase in confidence that there was no emerging trend and facilitates subsequent evaluation of intervention effects.

Occasionally, an investigator may obtain an extreme data point during baseline in the opposite direction of the change anticipated with treatment. This extreme point may be interpreted as suggesting that if there is any trend, it is in the opposite direction of treatment effects. Thus, investigators often shift phases when an extreme data point is noted in the previous phase in the direction opposite from predicted effects of the next phase. Yet, extreme scores in one direction are likely to be followed by scores that revert in the di-

rection of the mean. Thus, if the observation indicates a particularly high score on one occasion, the next occasion is likely to show a less extreme score. This characteristic is known as *statistical regression* and is a function of the correlation between consecutive data points.¹

It is important to be aware of the possibility of regression. Because the extreme score in one direction is likely to be followed by a much less extreme score that reverts toward the mean, it is unwise to shift phases following highly extreme scores. Shifting phases at this point would capitalize on regression. This improvement in performance would appear to be a function of shifting from baseline to intervention phases. However, the improvement might be a function of regression. As data continue to be collected in the new phase, the investigator could, of course, see whether the intervention is having an effect on behavior. However, if mean levels of performance are compared across phases, shifting phases at points of extreme scores could systematically bias the average performance in each of the phases. This could influence the conclusions that are drawn, especially if the phases are relatively brief.

Currently, there are no agreed upon objective decision rules for alternating phases in single-case experimental designs. Occasionally, investigators have attempted to specify objective criteria in advance that indicate the conditions under which baseline and experimental conditions will be alternated (e.g., Scott et al., 1973; Wincze, Leitenberg, & Agras, 1972). For example, phases can be shifted when variability about the mean level of performance falls within a specific range for a period of 5 days or when a given number of consecutive data points are not in the same direction above (or below) the mean. Attempts to develop an objective preset rule for an important design decision that has remained primarily subjective are laudible. However, the rules have not been sufficiently developed to this point to routinely exclude emerging trends or the influence of regression due to extreme scores in the data.

Checking the Intervention

In single-case experiments, the investigator manipulates specific conditions and assesses

their effects on behavior. If the dependent measure reflects change and the replication requirements within the design are met, the investigator is likely to assume that the intervention was responsible for behavior change. In most behavior modification interventions, in which single-case experimental designs are used with the greatest frequency, the intervention consists of the behavior of individuals (e.g., parents, teachers, peers) who interact with the client (e.g., child). For example, parental attention can be used to alter the behaviors of conduct problem children at home. When the intervention consists of behaviors in natural situations, the investigator has no assurance that the intervention is carried out correctly, or indeed is even implemented at all. This situation is to be contrasted with laboratory experiments in which the intervention can be implemented by reading instructions or playing an audio or video tape recorder and in which standardization of the intervention can be more readily assured.

When the behavior of individuals in contact with the client is the independent variable, it is especially important to gather data to ensure that the intervention is carried out. This is essential for at least three reasons. First, the behaviors of individuals who interact with the clients in naturalistic settings often are difficult to change (e.g., Breyer & Allen, 1975). Thus, an investigator cannot automatically assume that behavior-change agents will engage in the desired behaviors after receiving instructions or reminders. Second, when changes are made in behavior-change agents, they often are short-lived (e.g., Kazdin & Moyer, 1976). There is no assurance that once the intervention is implemented it will continue to be carried out in the appropriate fashion. Finally, individuals responsible for altering client behavior may show changes in their behaviors that extend

¹ The lower the correlation between consecutive data points, the greater the regression toward the mean. The reason for this is that the lower correlation between consecutive data points, the greater the amount of error in the scores. The random error that is likely to be present in an extreme score on one occasion is not likely to be present on the next occasion. Thus, a score that departs considerably from the overall mean is likely to be followed by a much less extreme score.

beyond the confines of the intervention (Chadwick & Day, 1971; Trudel, Boisvert, Maruca, & Leroux, 1974).

Even if experimental results meet the design requirements and are replicated across phases, precision can be increased by supplementary data showing that the intervention was carried out in the desired fashion. Without these data, the systematic changes in client behavior may be the result of some other influence in the setting. Indeed, it is quite possible given existing evidence on the performance of behavior-change agents that some aspect of their behavior other than the intended intervention led to change (Kazdin, 1977d).

Often it is desirable to have information about variables that may covary with the intervention and could plausibly account for change. For example, verbal approval and disapproval have proven to be very potent variables in altering child behavior. In programs evaluating other interventions (e.g., token reinforcement, nonverbal approval), it is essential to collect data on verbal behaviors of the behavior-change agents as well to assess whether they have covaried with the intervention (e.g., Kaufman & O'Leary, 1972; Kazdin & Klock, 1973). This assessment is important in situations in which behavior change might result from events that already have been shown to influence behavior in previous research and may be more plausible as influences than the variable of direct interest.

Comparing Alternative Interventions

Investigators using single-case designs frequently are interested in assessing the relative efficacy of different treatments with a given client. Comparing different treatments within a single set of subjects or a single subject is difficult because of the likely confound of treatments with sequence or order effects. It is still common to see investigators compare two or more different interventions in a version of an ABAB design. As a simple case, the design may be represented as an ABCA design (where A refers to baseline and B and C refer to alternative treatments). If performance levels of the client differ under the different treatments, the investigator can make statements about the relative efficacy of the

treatments. However, two treatments invariably are confounded with the order of appearance. One cannot interpret whether the first (or second) intervention led to more (or less) behavior change because of the treatment itself or because of the sequence in which it appeared.

A common technique is to add a return-to-baseline condition (an A phase) between the two interventions to show that original baseline level of performance is recovered (an ABACA design). Even if baseline levels are the same prior to each (B,C) intervention, this still does not correct for the confound of sequence effects. Merely because baseline phases prior to each intervention were equal in the level of performance does not mean that they are comparable in all other respects. Behavior change is not equally difficult to achieve on repeated occasions in which a treatment is presented and withdrawn. Additional interventions may need to be more or less effective than prior ones to effect change.

Investigators often conduct a more complex version of the design in which the effects of treatment are replicated at different points in the design for the subject. For example, in an ABCABC design, each treatment (B,C) is given twice. Even here, any conclusions reached about treatments are restricted to the particular order of treatment. The efficacy of C may result from the fact that it has followed B. The effects of C when presented without this prior history may be completely different. In short, the fixed sequence of treatments may lead to multiple-treatment interference (see Campbell & Stanley, 1963). Conclusions about the treatments must be tempered by acknowledging the possible contribution of one treatment on later treatments.²

The main problem with sequence effects in single-case experimental designs stems from the fact that investigators frequently rely on versions of the ABAB design. If this design

² Although the discussion of sequence effects may seem of academic interest, this is not the case. Both laboratory animal investigations (Grice & Hunter, 1964) and applications with clinical problems (White, Nielson, & Johnson, 1972) have shown that the effects of a given intervention may be affected by the treatments presented immediately before.

were used to compare treatments, more than one subject would be needed to ensure that the treatments are counterbalanced. Only if the treatments are presented in a different sequence can the investigator be assured that the interventions are differentially effective independent of the order in which they appear. Actually, using more than one subject is not a very useful solution. Unlike group designs in which the different sequences are replicated across several subjects, a single-case design with only two subjects would not allow separation of the sequence effects from unique subject characteristics. That is, if two treatments produce different effects across two subjects, the investigator cannot determine whether the order was important or whether the subjects merely responded differently to the treatments.

If the investigator is interested in single-case designs for comparing treatments, the ABAB design and its variations probably should be avoided. There are alternative single-case experimental designs suited to comparisons of different treatments. These include the multiple-element baseline or multiple-schedule design (Leitenberg, 1973; Ullman and Sulzer-Azaroff, 1975) or the simultaneous-treatment design (Kazdin, 1977c). In these designs, the separate interventions can be presented to an individual subject in such a way that the interventions are not confounded by sequence effects. Specifically, separate interventions are presented concurrently but can be balanced across conditions of administration.

Interobserver Agreement

In most single-case experimental designs, overt client behavior is assessed daily over the course of baseline and intervention conditions. In the majority of cases, observers record whether behavior has occurred based on their judgment and the definition of the response as originally specified. Occasionally, automated recording devices (e.g., time clock, plethysmograph), equipment requiring little or no judgment (e.g., scale), or devices that make scoring merely a clerical task (e.g., key to a paper-and-pencil test) are used, and there is little question whether observations are con-

sistent and accurate. In other cases, in which human judgment may figure more prominently in scoring the response, doubt can be raised about the adequacy of the assessment procedure and the consistency with which observations are made. In these latter cases, it is essential to assess the extent to which observers are collecting data consistently and accurately. Actually, assessment of client behavior and interobserver agreement might be viewed as purely measurement problems and be distinguished from design issues. However, single-case designs rely on behavioral observations, and assessment issues directly relate to evaluation of intervention effects.³

Interobserver agreement assesses whether the client's behavior is observed consistently. With low agreement, the data may differ greatly depending on who is scoring the behavior. Variation in the data due to the observer adds to any variability in client behavior and obscures actual performance. As noted earlier, evaluation of the intervention depends on obtaining data that have relatively little variability. Measurement error contributes to variability and makes subsequent evaluation more difficult. Indeed, if variability due to assessment error is extremely large, establishing a stable rate of behavior and evaluating the intervention may be impossible.

Evidence suggests that observers may stray from original behavioral codes in scoring client behavior (Kent & Foster, 1977). To ensure sustained accuracy and consistency in observer assessment, it is useful to check each observer against an independent standard and calculate the amount of agreement. Aside from a general trend to deviate from the original behavioral codes, observers may develop idiosyncratic coding tendencies so that they may

³ The assessment of interobserver agreement in single-case experiments has been a topic of major discussion in recent years. Basic issues are still actively discussed, including the manner in which interobserver agreement should be assessed, the conditions under which such methods are appropriate, the sources of bias that obscure interpretation of agreement data, and similar factors. These issues have been described in several publications (e.g., Hawkins & Dotson, 1975; Kent & Foster, 1977) including an invited series of articles in the *Journal of Applied Behavior Analysis* (1977, 10[4]).

see behaviors different from other observers. Periodic checks on interobserver agreement may help evaluate whether changes in scoring are occurring and whether particular observers are responsible for any of these changes.

Actually, research on interobserver agreement has revealed several factors that need to be taken into account when obtaining checks on observers (see Kazdin, 1977a; Kent & Foster, 1977). Observer awareness that agreement is being checked, who the other observer is, and reactions of the experimenter to the data that are reported have influenced interobserver agreement and the nature of the data obtained. As a general rule, a minimal condition for observation of overt behavior is the assessment of interobserver agreement at different points throughout the experiment. Agreement checks need to be dispersed over each of the phases to ensure that observer biases and changes in criteria for scoring responses are not likely to be confounded with treatment. Ideally, interobserver agreement should be assessed unobtrusively, because observer awareness that agreement is being checked influences the data. However, experimental arrangements required to accomplish unobtrusive agreement checks are difficult to meet (Kazdin, 1977a).

Issues in Data Interpretation

A major issue in evaluating treatment effects in research is the internal validity of experimental findings (Campbell & Stanley, 1963). Internal validity refers to assessing whether the intervention accounted for the results and essentially pertains to experimental design considerations that rule out potential sources of confound. The issue of internal validity is central to all research. However, single-case designs differ from traditional research in the criteria for determining whether an intervention accounted for the results. Failure to understand the criteria for evaluating results constitutes a major methodological issue.

Criteria for Evaluating Change

Single-case experimental designs attempt to structure the situation so that the effects of

the intervention and extraneous variables can be distinguished. The replication of intervention effects over time or across baselines is designed to make implausible the influence of various threats to internal validity. Just as the designs depart from traditional group experimentation, so do the criteria for evaluating the effectiveness of an intervention on behavior. Two criteria have been used to evaluate treatment effects in single-case experiments, namely, the experimental criterion and the therapeutic criterion (Kazdin, 1978b; Risley, 1970).

Experimental criterion. The experimental criterion refers to a comparison of behavior during the intervention with what it would be like if the intervention had not been implemented. This criterion, of course, characterizes group research designs as well. However, single-case experiments usually do not fulfill this criterion by appealing to statistical analyses. Rather, the experimental criterion is met by replicating treatment effects over time. Repeatedly showing that alteration of contingency changes the level of performance in relation to the previous phase fulfills the experimental criterion. The strength of the demonstration derives from showing that performance during a given phase violates the predicted level of performance of the prior phase before the intervention was introduced. Each design accomplishes the replication of treatment effects in a different way. For example, in the ABAB design, the intervention is replicated over time for a single subject or group of subjects. Similarly, in a multiple-baseline design, the effect is replicated across separate baselines.

In practice, whether the results meet the experimental criterion is determined in various ways. First, if performance during an intervention phase does not overlap with performance during the baseline phase when these data points are plotted over time, the effects usually are regarded as reliable. The replication of nonoverlapping distributions during different treatment phases strongly argues for the effects of treatment. Second, a more typical criterion for experimental evaluation is related to the trends in each phase. If baseline and intervention conditions show changes in trends as the phases are alternated, the re-

liability of the data and effects of the intervention usually are inferred.

One of the problems with experimental evaluation of single-case method designs is that they have relied almost exclusively on visual inspection to determine whether the magnitude of change across phases is significant. The problem with visual inspection is that individuals who peruse the data may not see eye to eye. For example, Jones, Weinrott, and Vaught (Note 1) demonstrated considerable disagreement among judges asked to visually inspect the data from the same single-case experimental design.

Recently, statistical evaluation of single-case data has received increased attention (Kazdin, 1976; Michael, 1974). Such evaluations are likely to be useful in situations in which the ideal conditions such as stable baselines and relatively small variability cannot be met. Thus, the reliability of treatment effects can be examined in situations in which visual inspection might be weak or unreliable. However, the use of statistics with single-case designs is controversial (Michael, 1974). Also, statistical tests appropriate for single-case analyses are less familiar to most investigators because they are not the ones commonly taught in graduate statistics in psychology. Further, the tests include relatively restrictive assumptions that may dictate selection of the design and the manner in which treatment is implemented (see Kazdin, 1976). Because of the infrequent use of statistics in single-case experiments and lack of familiarity with the statistical properties of the tests, it is no surprise that inappropriate applications of statistical tests already have entered the literature. In any case, statistical tests are available for evaluating the reliability of intervention effects in single-case experimental designs. These tests have been illustrated at length elsewhere (Jones et al., 1977; Kazdin, 1976).

Therapeutic criterion. The therapeutic criterion for evaluating single-case experimental designs refers to the value or importance of behavior change, that is, the clinical significance of treatment effects. This criterion refers to whether the extent of behavior change achieved during treatment enhances the client's functioning in everyday situations

(Risley, 1970). Implicit in this criterion, of course, is that the behavior selected for alteration is itself of clinical or social importance.

The therapeutic criterion is more difficult to satisfy than is the experimental criterion. Showing a reliable treatment effect, through visual inspection or statistical analysis, has no necessary bearing on the importance of the change. For example, a single-case demonstration might well show a reduction of self-destructive behavior from 100 to 50 instances per hour. Even though the reduction is relatively large and might meet the most stringent statistical test, the remaining level of behavior is far removed from the "normal social interaction" to which the individual might someday return. To be of unequivocal therapeutic significance, the intervention would need to eliminate self-destructive behavior.

The ease of evaluating the importance of clinical change in the above example derives from the fact that intense destructive behavior is maladaptive whenever it occurs. Yet, for many behaviors, the level and intensity of behavior rather than merely its presence or absence dictate whether it is acceptable. This makes satisfying the therapeutic criterion much more difficult. Indeed, the precise criterion for evaluating whether the effects of treatment are clinically important is difficult to specify. Part of the reason is that individuals in everyday life (e.g., parents, teachers, friends) and the individual himself or herself determine the level of behavior that is acceptable or deviant.

Recently, single-case experimental methodology has made efforts to assess objectively whether the effects of treatment are clinically significant. The procedures for assessing the clinical importance of treatment effects are referred to as *social validation* (Wolf, Note 2) and have been incorporated into an increasing number of investigations (Kazdin, 1977b). Generally, social validation of treatment effects consists of determining whether behavior changes are clinically important in the social context in which the client functions. This is accomplished in one of two ways, referred to as social comparison and subjective evaluation methods (Kazdin,

1977b). With the *social comparison method*, the behavior of the client before and after treatment is compared with the behavior of "nondeviant" peers. The question asked by this comparison is whether the client's behavior after treatment is distinguishable from the normative range of behavior of his or her peers. With the *subjective evaluation method*, behavior is evaluated by individuals who are likely to have contact with the client to determine whether the change made during treatment is important. The question addressed by this method is whether behavior changes demonstrated in treatment lead to qualitative differences in how the client is viewed by others.

In using the social comparison method to evaluate treatment, it is essential to identify individuals who are similar to the client in subject and demographic variables but who have been identified as showing acceptable behavior. Presumably, peers of the client identified in such a fashion should differ markedly from the client in the target behavior prior to treatment. After treatment, the behavior of the client and peers can be compared again. If treatment has effected clinically significant change, this should be demonstrated by showing that the client's behavior has moved to within acceptable, that is, "normative," levels of functioning. Several studies have shown that treatment effects bring initially deviant levels of behavior within normative levels for social interaction or conduct problems of children (e.g., O'Connor, 1972), basic reading skills of the mentally retarded (Azrin & Armstrong, 1973), locus of control and self-concepts of delinquents (Eitzen, 1975), social skills of unassertive college students (e.g., McFall & Twentyman, 1973), and others (see Kazdin, 1977b, for a review).

In applying the subjective evaluation method, individuals who normally interact with the client or who are in a special position (e.g., through expertise) to judge a particular behavior provide global evaluations to assess how well the client is functioning. Essentially, the global appraisal is used to evaluate whether the specific behaviors changed have implications for how the client is viewed by others. Obviously, the global ratings are used only to supplement objective behavioral

data, since studies have demonstrated that global judgments are much more likely to reflect bias than are specific objective measures (Kazdin, 1973; Kent, O'Leary, Diamant, & Dietz, 1974; Schnelle, 1974).

The subjective evaluation method can be used both to help identify important behaviors to assess and focus on in treatment as well as to evaluate qualitative improvements associated with behavior change. Studies have used global assessments prior to treatment to determine whether the specific behaviors that are observed and included in treatment are relevant to individuals in contact with the client in everyday situations (e.g., Werner et al., 1975). More commonly, subjective evaluations are used to evaluate pretreatment and posttreatment levels of behavior. Several studies have shown that behavior changes have improved qualitative global judgments of individuals in contact with the clients for programs training delinquent girls how to communicate appropriately with others (e.g., Maloney et al., 1976), developing public-speaking skills in adults (Fawcett & Miller, 1975), training problem-solving skills of lower socioeconomic adults in positions that may influence decision making (Briscoe, Hoffman, & Bailey, 1975), and altering performance of conduct problem children (Kent & O'Leary, 1976).

Social comparison and subjective evaluation methods are not without problems. For example, identifying appropriate peer groups to which client behavior can be compared, deciding whether normative levels are the appropriate standard or themselves should be subject to change, and determining precisely what global judgments actually mean are only some of the problems for evaluating the therapeutic criterion (Kazdin, 1977b). However imperfect, the methods show considerable promise in assessing the clinical importance of behavior change and need to be incorporated routinely into clinical research. Although the bulk of studies using social validation procedures relied on single-case experimental designs, many of the studies were traditional between-group designs (e.g., Kent & O'Leary, 1976; McFall & Twentyman, 1973). The validation procedures are not restricted to any particular design and are emphasized here

only because single-case experimental research in behavior modification has embraced a therapeutic criterion as a central feature of treatment evaluation.

Generality of the Results

An important issue, often noted in the form of an objection to single-case experimental designs, is that the results from an individual case may not necessarily apply or be generalizable to other subjects.⁴ The generality of experimental findings in between-group and single-case research is a weighty topic that cannot be developed fully here (see Hersen & Barlow, 1976; Kazdin, 1973). Suffice it to say that mere numbers of subjects do not, in their own right, guarantee generality of the results. Indeed, the vast majority of between-group investigations evaluates results on the basis of average (mean) group performance, which does not provide insights about the generality of a particular effect within a given experimental condition. Yet, this latter problem is a function of not looking at all of the individual subject data within a between-group investigation. With single-case demonstrations, there is no immediate possibility to assess generality, by definition.

Although it is quite possible that the effects of treatment demonstrated with the individual case may not generalize to others, this does not appear to be a problem in contemporary research. Findings obtained in single-case demonstrations appear to be as generalizable, if not much more so, than demonstrations using other designs. The reason for this does not seem to be related to the designs per se but rather to the type of interventions that are commonly evaluated. Investigators who use single-case designs have made a special point to look for interventions that produce dramatic changes in behavior. Interventions with such effects for the single case are likely to generalize more broadly than are interventions that meet the relatively weaker criterion of statistical significance based on group averages that characterize between-group research.

Although generality of the results in single-case designs is not an inherent problem, investigation of the dimensions along which

generality might occur is difficult. The designs characteristically are weak in revealing subject characteristics that may interact with a specific treatment. Focusing on one subject does not allow for the systematic comparison of different treatments across multiple subjects who differ in various characteristics, at least within a single experiment. Examination of subject variables is more readily accomplished by between-group research, specifically, factorial designs in which intervention effects can be analyzed separately according to characteristics of the subjects. Yet, single-case experiments have focused on developing potent interventions that tend not to depend heavily on subject characteristics. This has been a matter of philosophy rather than a simple design consideration (Kazdin, 1978a). The ultimate test of generality of findings among subjects or any other condition is replication. Single-case experiments commonly replicate intervention effects across a wide range of clients (e.g., Kazdin, 1977d).

Concluding Comments

Single-case experimental designs offer unique advantages for research in clinical psychology because they provide an empirical and scientific basis for investigating interventions with individual clients. The diverse designs permit examination of different outcome questions, such as the effects of an overall treatment package, analysis of components of a package, and comparison of different treatments. Single-case designs make available areas not easily studied in group research. For example, many clinical problems cannot be studied on a group basis simply because few clients come for treatment or the behaviors are extremely rare.

Single-case designs make demands on an investigator that may not be easily met in many clinical situations. Initially, client behavior needs to be observed continuously, a task

⁴ It is only fair to point out that many "single-case" designs utilize more than one individual; indeed, large groups often are included in such designs. Also, some design variations including one of the multiple-baseline designs routinely rely on more than one individual.

more difficult to implement than traditional pretreatment and posttreatment assessment. Second, behavioral measures usually are constructed to assess the individual client's problem. Third, some designs require reinstating conditions that compete with effecting durable therapeutic changes. Finally, many important questions are not easily addressed, such as the interaction of client, therapist, or setting variables with the intervention. Interactions of this sort are more readily addressed in group factorial designs.

Single-case designs appear relatively straightforward, a characteristic that may foster rather than ameliorate the methodological and interpretive problems highlighted in this article. Complexities of the designs derive from the lack of clear rules for deciding how long to gather data, when to implement treatment, and the criteria for evaluating change. Quantitative guides for making decisions about changing from baseline to intervention phases and for evaluating the statistical reliability of interventions are available but are still infrequently used in the literature. The best guide for appropriate use and interpretation of single-case experiments is recognition of the basic logic of the designs and common methodological problems. The present article addressed some of the major problems that impede interpretation of findings with single-case experiments.

Reference Notes

1. Jones, R. R., Weinrott, M., & Vaught, R. S. *Visual vs. statistical inference in operant research*. Paper presented at the annual meeting of the American Psychological Association, Chicago, September 1975.
2. Wolf, M. M. *Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C., September 1976.

References

1. Ayllon, T., & Haughton, E. Modification of symptomatic verbal behaviour of mental patients. *Behaviour Research and Therapy*, 1964, 2, 87-97.
2. Martin, N. H., & Armstrong, P. M. The "mini-meal"—A method for teaching eating skills to the profoundly retarded. *Mental Retardation*, 1973, 11, 9-13.
3. Baer, D. M., Wolf, M. M., & Risley, T. R. Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1968, 1, 91-97.
4. Breyer, N. L., & Allen, G. J. Effects of implementing a token economy on teacher attending behavior. *Journal of Applied Behavior Analysis*, 1975, 8, 373-380.
5. Briscoe, R. V., Hoffman, D. B., & Bailey, J. S. Behavioral community psychology: Training a community board to problem solve. *Journal of Applied Behavior Analysis*, 1975, 8, 157-168.
6. Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally, 1963.
7. Chadwick, B. A., & Day, R. C. Systematic reinforcement: Academic performance of under-achieving students. *Journal of Applied Behavior Analysis*, 1971, 4, 311-319.
8. Eitzen, D. S. The effects of behavior modification on the attitude of delinquents. *Behaviour Research and Therapy*, 1975, 13, 295-299.
9. Fawcett, S. B., & Miller, L. K. Training public-speaking behavior: An experimental analysis and social validation. *Journal of Applied Behavior Analysis*, 1975, 8, 125-135.
10. Glass, G. V., Willson, V. L., & Gottman, J. M. *Design and analysis of time-series experiments*. Boulder: Colorado Associated University Press, 1974.
11. Grice, C. R., & Hunter, J. J. Stimulus intensity effects depend upon the type of experimental design. *Psychological Review*, 1964, 71, 247-256.
12. Hartmann, D. P., & Hall, R. V. The changing criterion design. *Journal of Applied Behavior Analysis*, 1976, 9, 527-532.
13. Hawkins, R. P., & Dotson, V. A. Reliability scores that delude: An Alice in Wonderland trip through the misleading characteristics of interobserver agreement scores in interval recording. In E. Ramp & G. Semb (Eds.), *Behavior analysis: Areas of research and application*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.
14. Hersen, M., & Barlow, D. H. *Single-case experimental designs: Strategies for studying behavior change*. New York: Pergamon Press, 1976.
15. Jones, R. R., Vaught, R. S., & Weinrott, M. Time-series analysis in operant research. *Journal of Applied Behavior Analysis*, 1977, 10, 151-166.
16. Kaufman, K. F., & O'Leary, K. D. Reward, cost and self-evaluation procedures for disruptive adolescents in a psychiatric hospital. *Journal of Applied Behavior Analysis*, 1972, 5, 293-309.
17. Kazdin, A. E. Role of instructions and reinforcement in behavior changes in token reinforcement programs. *Journal of Educational Psychology*, 1973, 64, 63-71.
18. Kazdin, A. E. Statistical analyses for single-case experimental designs. In M. Hersen & D. H. Barlow, *Single-case experimental designs: Strategies for studying behavior change*. New York: Pergamon Press, 1976.

- Kazdin, A. E. Artifact, bias, and complexity of assessment: The ABC's of reliability. *Journal of Applied Behavior Analysis*, 1977, 10, 141-150. (a)
- Kazdin, A. E. Assessing the clinical or applied significance of behavior change through social validation. *Behavior Modification*, 1977, 1, 427-452. (b)
- Kazdin, A. E. The influence of behavior preceding a reinforced response on behavior change in the classroom. *Journal of Applied Behavior Analysis*, 1977, 10, 299-310. (c)
- Kazdin, A. E. *The token economy: A review and evaluation*. New York: Plenum Press, 1977. (d)
- Kazdin, A. E. *History of behavior modification: Experimental foundations of contemporary research*. Baltimore, Md.: University Park Press, 1978. (a)
- Kazdin, A. E. Methodology of applied behavior analysis. In T. Brigham & A. C. Catania (Eds.), *Handbook of applied behavior research: Social and instructional processes*. New York: Irvington, 1978. (b)
- Kazdin, A. E., & Klock, J. The effect of nonverbal teacher approval on student attentive behavior. *Journal of Applied Behavior Analysis*, 1973, 6, 643-654.
- Kazdin, A. E., & Moyer, W. Training teachers to use behavior modification. In S. Yen & R. McIntire (Eds.), *Teaching behavior modification*. Kalamazoo, Mich.: Behaviordelia, 1976.
- Kent, R. N., & Foster, S. L. Direct observational procedures: Methodological issues in naturalistic settings. In A. R. Cimminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Kent, R. N., & O'Leary, K. D. A controlled evaluation of behavior modification with conduct problem children. *Journal of Consulting and Clinical Psychology*, 1976, 44, 586-596.
- Kent, R. N., O'Leary, K. D., Diamant, C., & Dietz, A. Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology*, 1974, 42, 774-780.
- Leitenberg, H. The use of single case methodology in psychotherapy research. *Journal of Abnormal Psychology*, 1973, 82, 87-101.
- Maloney, D. M., et al. Teaching conversation-related skills to predelinquent girls. *Journal of Applied Behavior Analysis*, 1976, 9, 371.
- McFall, R. M., & Twentyman, C. T. Four experiments on the relative contributions of rehearsal, modeling, and coaching to assertion training. *Journal of Abnormal Psychology*, 1973, 81, 199-218.
- Michael, J. Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis*, 1974, 7, 647-653.
- O'Connor, R. D. Relative efficacy of modeling, shaping, and the combined procedures for modification of social withdrawal. *Journal of Abnormal Psychology*, 1972, 79, 327-334.
- Risley, T. R. Behavior modification: An experimental-therapeutic endeavor. In L. S. Hamerlynck, P. O. Davidson, & L. E. Acker (Eds.), *Behavior modification and ideal mental health services*. Calgary, Alberta, Canada: University of Calgary Press, 1970.
- Schnelle, J. F. A brief report on invalidity of parent evaluations of behavior change. *Journal of Applied Behavior Analysis*, 1974, 7, 341-343.
- Scott, R. W., Peters, R. D., Gillespie, W. J., Blanchard, E. B., Edmunson, E. D., & Young, L. D. The use of shaping and reinforcement in the operant acceleration and deceleration of heart rate. *Behaviour Research and Therapy*, 1973, 11, 179-185.
- Trudel, G., Boisvert, J., Maruca, F., & Leroux, P. Unprogrammed reinforcement of patients' behaviors in wards with and without token economy. *Journal of Behavior Therapy and Experimental Psychiatry*, 1974, 5, 147-149.
- Ulman, J. D., & Sulzer-Azaroff, B. Multielement baseline design in educational research. In E. Ramp & G. Semb (Eds.), *Behavior analysis: Areas of research and application*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.
- Werner, J. S., et al. "Intervention package": An analysis to prepare juvenile delinquents for encounters with police officers. *Criminal Justice and Behavior*, 1975, 2, 55-83.
- White, G. D., Nielson, G., & Johnson, S. M. Time-out duration and the suppression of deviant behavior in children. *Journal of Applied Behavior Analysis*, 1972, 5, 111-120.
- Wincze, J. P., Leitenberg, H., & Agras, W. S. The effects of token reinforcement and feedback on the delusional verbal behavior of chronic paranoid schizophrenics. *Journal of Applied Behavior Analysis*, 1972, 5, 247-262.

Received August 1, 1977 ■

Stimulus Sampling in Clinical Research: Representative Design Reviewed

Brendan A. Maher
Harvard University

Brunswik's concept of representative design is reviewed with special reference to studies of clinical bias. The limitations of single-stimulus, actor-script, and serial replications are discussed. No satisfactory alternatives exist to adequate sampling of stimulus persons.

More than 30 years ago, Egon Brunswik (1947) pointed out that if we wish to generalize the results of a psychological experiment to populations of subjects and to populations of stimuli, we must sample from both populations. This argument was elaborated by him in other articles and was summarized cogently in a short article by Hammond (1948). The purpose of this article is to review the issues that Brunswik raised and to examine some of their implications for contemporary research in clinical psychology.

Brunswik's thesis is very simple. When we conduct an experiment intended to investigate the effect of different values of an independent variable on a population, we always take care to draw a sample of subjects that is representative of the population in question. We do so, naturally, because we recognize the range of variation that exists in populations of individuals. We wish to make sure that deviant individual values do not distort our estimate of the parameters of the population. If the stimuli that we use are defined in physical units, we are (or should be) careful to confine our generalizations to the range of values actually included in the study. When physical units are involved, we have relative confidence that the stimulus can be replicated by another investigator, provided that the detailed descrip-

tion of the stimulus is followed carefully. Should a subsequent investigator change one or more of these attributes, we are not surprised if there is a concomitant change in the responses that are made to the stimulus.

When the stimuli to which the subjects respond cannot be defined in physical units and are likely to vary within a population, a different situation arises. Outstanding examples are to be seen in research directed to the investigation of the effects of human beings as stimuli that elicit behavior from other human beings. Consider some instances drawn from recent volumes of this journal. Acosta and Sheehan (1976) reported that Mexican American subjects viewed an Anglo American professional therapist as more competent than a Mexican American professional when all other variables were matched. Babad, Mann, and Mar-Hayim (1975) reported that trainee clinicians who were told that a testee was a high-achieving upper-middle class child assigned higher scores to Wechsler Intelligence Scale for Children (WISC) responses than did another sample of clinicians who were led to believe that the same responses had been made by an underachieving deprived child. Research of this kind is generally cast in terms of a hypothesis that members of a specified population respond in discriminatory fashion to members of certain other populations. Thus, for example, we encounter such questions as, Do physicians give less adequate medical care to ex-mental patients than they do to normal medical patients? (Farina, Hagelauer, & Holzberg, 1976) and Are thera-

My appreciation is due to Winifred Barbara Maher for her critical reading of early drafts of this article. Requests for reprints should be sent to Brendan A. Maher, Department of Psychology and Social Relations, Harvard University, Cambridge, Massachusetts 02138.

pists with a behavioral orientation less affected by the label *patient* when evaluating observed behavior than are therapists of psychodynamic persuasion? (Langer & Abelson, 1974).

Single-Stimulus Design

Human attributes are generally distributed in such a fashion that any one of them is likely to be found in conjunction with a wide variety of others. Let us consider an investigation of bias toward ex-mental patients. To belabor the obvious a little, we can note that the attribute *ex-mental patient* can be associated with any measure of intelligence, age, sex, education, socioeconomic status, physical attractiveness, and so forth. It is true that some of these attributes may have significant correlations with each other; a patient of upper socioeconomic status is quite likely to have had substantial education, for example. Nonetheless, even the largest of these correlations is quite modest, and the population of ex-mental patients to which we wish to generalize will have a wide range of values on these attributes.

When we employ only one person as a stimulus, we are faced with the fact that the specific values of some of the other attributes possessed by this person will also have stimulus value that will be unknown and uncontrolled. Responses made by a sample of the normal population to an ex-mental patient who is female, young, attractive, articulate, and intelligent may well be different from those made to a normal control who is male, old, ugly, incoherent, and dull. These differences cannot be assigned to the patient/non-patient status of the two stimulus persons, as many other unidentified differences were uncontrolled. At first sight it may appear that this problem is solved by the simple expedient of matching the patient and the control on all variables other than that of patient status. Unfortunately, this can only be achieved at the cost of further difficulties. We do not know the full range of variables that should be matched, and hence this solution necessarily involves resort to an actor and a script, barring the unlikely availability of discordant monozygotic twins for research purposes! Scripts bring with them some special prob-

lems, of which more will be said later. The main point to note here is that the use of a *single* human stimulus acting as his or her own control fails to deal with the problem of the *interaction* of the attribute under investigation with those that have been controlled by matching. Pursuing for a moment the example of responses to the label *ex-mental patient*, let us consider a hypothetical study using a male actor with athletic physique and vigorous movements. The willingness of a normal subject to accept this individual as a fellow worker, neighbor, or friend may well be influenced by the perception that the ex-patient, if violent, could be dangerous. Had the actor been older and visibly frail, the reaction might well be different. Under the first set of circumstances, the bias hypothesis would probably be confirmed, and under the second set, the null hypothesis might fail to be rejected.

An additional difficulty is incurred by the single-stimulus own-control strategy. We cannot determine whether a finding of no difference between group means is due to the weakness of the hypothesis, errors of method, or the inadvertent selection of an atypical stimulus person to represent one or both conditions. An example of the complexities of interpretation with this design can be found in Farina et al. (1976). These investigators hypothesized that physicians would provide less adequate medical care to former mental patients than to normal medical students. To test this hypothesis one stimulus person, a 23-year-old male graduate student, approached 32 medical practitioners. In each case he entered the doctor's office

carrying a motorcycle helmet and a small knapsack. . . . The same symptoms were reported to all doctors. Stomach pains suggestive of ulcers were selected to be neither clearly psychiatric nor unrelated to the mind. . . . Every other practitioner was told the pains had first occurred 9 months earlier while the patient was traveling around the country. The remaining 16 doctors were also informed that the pains had appeared 9 months earlier, but at that time the patient reported being in a mental hospital. (Farina et al., 1976, p. 499)

No significant difference of any relevance was found in the kind of medical care given by the practitioners under either condition. In

conclusion, the authors stated that "a former mental patient seems to receive the same medical treatment as anyone else" (p. 499).

Logically, several conclusions are compatible with this finding. One obviously valid conclusion is that a young male motorcyclist with the symptoms of ulcers receives a certain class of treatment whether or not he describes himself as a former mental patient. We cannot tell whether this treatment is the same, better, or worse than that typically given to a random sample of the normal population of patients who seek treatment for stomach pains, as no such sample was obtained. A substantial number of physicians may have had opinions about motorcyclists as unfavorable as those that they were hypothesized to have about former mental patients, and hence both conditions produced equally inadequate medical care. Alternatively, the physicians may have felt the necessity to be unusually careful in providing care to individuals who might be assumed to be irresponsible (such as motorcyclists and mental patients), and hence they provided better than average care. Finally, medical practice may be sufficiently precise that the adequate procedures to follow with patients who complain of stomach pains that no real room for bias exists, the treatment provided being the same as would be given to any sample of patients.

We can summarize the limitations of the single-stimulus design as follows:

1. Obtained differences may be due to the validity of the tested hypothesis or to the effect of uncontrolled stimulus variables in statistical interaction with the intended independent variable. No method of distinguishing between these two explanations is possible.

2. Lack of difference may be due to the invalidity of the hypothesis, undiscovered methodological factors such as subject sampling error, or the presence of an uncontrolled stimulus variable operating to either counteract or raise this effect to a ceiling value in both experimental and control situations.

It is readily apparent that the problem of uncontrolled attributes occurring in a single stimulus person can only be solved by the provision of an adequate sample of stimulus persons, since they will tend to cancel each

other out. No satisfactory solution is possible within the single-stimulus design.

Scripts and Manticores

Some investigators have attempted to solve the problems of the single-subject stimulus by fabricating scripts without the use of a human actor to present them. Case histories, dossiers, vignettes, audiotapes, or other devices have been used to reduce the effects of the uncontrolled aspects of a human stimulus. Thus, in a study by Babad et al. (1975), the trainee clinicians were given only the WISC protocol and did not see the child who was alleged to have been tested. These manufactured materials may be termed *scripts*. Scripts may be taken from existing sources of genuine material, such as clinical files; they may be created de novo in accordance with prior theoretical guidelines or in an attempt to present an ideal "typical" case.

When the script is drawn from original clinical files, the investigator is assured that at least one such case exists in nature. The limitations on the results obtained from such scripts are, in principle, the same as those that plague any single-stimulus design. Some minor advantage accrues to the method, however, in that the number of uncontrolled accidental attributes has been reduced by the elimination of those attributes associated with physical appearance, dress, and so forth. When the script is fabricated for research purposes, a new problem develops—namely that in devising material according to theoretical guidelines, a case is created that like the manticore, may never have existed in nature. We can imagine a hypothetical investigation of the attitudes of males toward females of varying degrees of power. Varying naval ranks with male and female gender of the occupant of each rank, we create the dossier of an imaginary female Fleet Admiral. Whatever our male subject's response to this dossier may be, we have no way of knowing whether it is due to the theoretically important combination of high rank with female gender or to the singularity of a combination that is, as yet, unknown to human experience.

For a recent illustration of this problem, we can turn to Acosta and Sheehan (1976). They

presented groups of Mexican American and Anglo American undergraduates with a videotaped excerpt of enacted psychotherapy. Each group saw an identical tape, except that in one version the therapist spoke English with a slight Spanish accent and in the other version the accent was standard American English. Some subjects were told that the therapist was a highly trained professional; the others were told that the therapist was a paraprofessional of limited experience. There were thus four experimental conditions and two kinds of subjects. The Spanish-accent tape of a trained professional was introduced with a background vinette describing the therapist as American born of Mexican parentage and as having a Harvard doctorate in his field and a distinguished professional record. For the American-English-accent tape, the therapist was introduced with the same vinette but with an Anglo-Saxon name and parentage identified as Northern European. Anglo American ratings of the therapist's competence were uninfluenced by the ethnic identification, whereas Mexican Americans rated the Mexican American therapist less favorably than the Anglo American therapist.

In their discussion of this somewhat surprising result, the authors noted that the number of Mexican American therapists actually in practice in the United States shortly before the study was done was 48 (28 psychologists and 20 psychiatrists). We do not know what characteristics would be typical of this population, and no attempt seems to have been made to ascertain them before preparing the script. There is, therefore, no way to be sure that the therapeutic style, choice of words, gesture, and so forth, were authentically typical of actual Mexican American therapists. Given that essentially the same script was used for both ethnic conditions, we must conclude that either one or the other version of the script was ethnically inaccurate or, less likely, that the only actual difference that would be seen in the comparative behaviors of Mexican American and Anglo American therapists would be their accent. In brief, we cannot ignore the possibility that the Mexican American subjects disapproved of the Mexican American therapist not because he was Mexican American but because his behavior was not representative of that of actual Mexican

Americans. Like the woman admiral, he may have presented a combination of characteristics that is theoretically possible but unknown in the experience of the subjects responding to it. The only guarantee that a script is free from impossible or improbable combinations of variables is when it is directly drawn from an actual clinical case or other human transaction. We cannot produce a fictional script of a psychotherapeutic session with any confidence that it is as representative as a transcript of an actual session. The ideal or typical therapeutic interview may be as rare as the perfect textbook case of conversion hysteria or as a stereotypical Mexican American. This rarity or implausibility may well determine a subject's response far more than the attributes that were planned to make it appear typical. Our hesitation in generalizing from a single stimulus case to a population of cases is increased substantially by the prospect of generalizing from a case that is not known to have existed at all.

Representative Design

The moral to the foregoing review is simple. If we wish to generalize to populations of stimuli, we must sample from them. Only in this way can we be confident that the various attributes that are found in the population will be properly represented in the sample. Those attributes that are significantly correlated with membership in the population will appear in appropriate and better-than-chance proportions; those attributes that are uncorrelated with population membership will appear in chance proportions but will not affect the outcomes. If we intend to draw conclusions about the way in which physicians treat former mental patients, we must sample physicians and former mental patients. If we wish to know what Mexican American students think of Mexican American therapists, we must sample students and therapists. This is the essence of Brunswik's (1947) concept of representative design. There is no satisfactory alternative to it. Nonetheless, the use of representative design is rarely, if ever, seen in reported research. There are, in my opinion, three reasons for this. First, many clinical psychologists are unaware of Brunswik's work. The remedy for this is obvious and easy to

apply. Second, there is a common failure to understand that the replication of single-stimulus studies with additional single-stimulus studies cannot create accumulated representative design unless the selection of single-stimulus persons was achieved by sampling.

Let us consider a hypothetical series of studies of the effect of examiner gender on children's test responses. In the population of examiners, there are likely to be attributes that distinguish males from females in addition to those that are inseparable from gender. Thus the proportions of married and single persons, prior experience with children, knowledge of various hobbies, mean age, prior locale of undergraduate education, and so forth, may differ between the two groups. In the first study we use one male examiner and one female examiner, each with 1 year of experience. Using samples of male and female children, we find differences in test responses attributable to examiner gender. Conscious of the fact that we included inexperienced examiners, we replicate the study with one male examiner and one female examiner each with 3 years of experience. Now we find no difference. Our series ends when we have made gender comparisons for examiners with 1, 3, 5, 7, 9, 11, 13, 15, 17, and 19 years of experience. We found significant examiner effects at every level of experience except 3 and 5 years. As 8 of our 10 studies have found significant differences due to gender, we conclude that there is a generalizable finding. We might even treat the entire series as a single experiment comparing the group of 10 male examiners with the group of 10 female examiners and find a statistically significant difference between the mean test responses elicited by one group versus the other.

To accept this conclusion it is first necessary to know what the true proportion of the total population of examiners at each level of experience is. If the experience range of 3-5 years includes 65% of all examiners, our best conclusion is that gender differences have not been established. The reason is, of course, that the "sample" of examiners was not representative of the population to which it is intended to generalize, being underrepresented in the 3- to 5-year experience range. Note that we cannot handle this by some proportional weighing of the data obtained from the ex-

aminers with 3-5 years of experience, as the results obtained from those comparisons suffer from the limitations of single-stimulus design and might well be due to the effects of uncontrolled differences between examiners other than gender.

A third reason for the failure to use representative design is that it is laborious and expensive. Providing an adequate sample of stimulus persons, each of whom is to be observed by an adequate sample of subjects, necessarily involves large numbers and long hours. For some investigators it is, as one of my correspondents put it, "too hard to do it right."

There is, however, no satisfactory alternative to doing it right. Clinical psychology is concerned with real people and not with hypothetical collections of attributes. Our research into the behavior of patients, therapists, diagnosticians, normal persons, and the like, must produce generalizations that are valid for actual populations of these people. Conclusions based on inadequate sampling may be worse than no conclusions at all if we decide to base our clinical decisions on them. If the patience and time that it takes to do it right create better science, our gratitude should not be diminished by the probability that fewer publications will be produced.

References

- Acosta, F. X., & Sheehan, J. G. Preferences toward Mexican American and Anglo American psychotherapists. *Journal of Consulting and Clinical Psychology*, 1976, 44, 272-279.
- Babad, E. Y., Mann, M., & Mar-Hayim, M. Bias in scoring WISC subtests. *Journal of Consulting and Clinical Psychology*, 1975, 43, 268.
- Brunswik, E. *Systematic and representative design of psychological experiments*. Berkeley: University of California Press, 1947.
- Farina, A., Hagelauer, H. D., & Holzberg, J. D. Influence of psychiatric history on physician's response to a new patient. *Journal of Consulting and Clinical Psychology*, 1976, 44, 499.
- Hammond, K. Subject and object sampling—A note. *Psychological Bulletin*, 1948, 45, 530-533.
- Langer, E. J., & Abelson, R. P. A patient by any other name . . . : Clinician group difference in labeling bias. *Journal of Consulting and Clinical Psychology*, 1974, 42, 4-9.

Common Methodological Problems in Factor Analytic Studies

Andrew L. Comrey
University of California, Los Angeles

Investigators are urged to plan factor analytic studies prior to collecting the data, to formulate a hypothesized factor structure, to develop several relatively pure measures of each factor expected, and to select an appropriate sample of at least 200 cases. Continuous variables should be used rather than dichotomous variables wherever possible. Programmatic series of studies are preferred over one-shot investigations. Putting unities in the diagonals and rotating all factors with eigenvalues of one or more is discouraged, because this procedure tends to give communalities that are too high, produces too many factors, and distorts the rotational solution, especially when analytic rotational programs are used. In some situations, a computer-assisted hand rotational solution is most likely to give satisfactory results. Mathematical algorithms designed to approximate simple structure work well only in situations properly designed for their application.

Despite the fact that factor analysis has been roundly condemned by many behavioral scientists as being without redeeming social or scientific value, the number of factor analytic investigations reaching manuscript form appears to be on a geometric upward course. Why? For one thing, critics of factor analysis often are not sufficiently knowledgeable about factor analysis to discriminate between a proper and an improper use of the technique. They are ready, therefore, to generalize from a collection of poor studies to the conclusion that all factor analytic studies are worthless. There are situations in which benefits can be derived from the proper use of factor analysis, and this fact can account at least to some extent for its expanded use despite persistent and heavy criticism. Probably a greater reason for the increase in use, however, is the ready availability of programs at various computer centers that will produce factor analytic results cheaply. The temptation to use such a convenient resource in this publish-or-perish age is apparently overwhelming. If all else fails in dealing with a body of data, there is

always the computer and factor analysis standing ready to rescue an otherwise fruitless quest for publishable results. In the face of an avalanche of poor factor analytic studies, it is not surprising that many thoughtful scientists and journal editors have come to distrust the method. For those who are willing to take the trouble to learn how to use such methods properly, however, factor analysis can serve as a powerful analytic tool to aid the scientist in his search for reliable knowledge. The purpose of this article is to warn inexperienced users about some of the pitfalls that they may encounter in trying to use factor analytic methods in their research.

Design of Factor Analytic Investigations

In experimental studies it is rather unlikely that a statistical treatment procedure can be found after the data have been collected that will be entirely satisfactory unless some thought was given to the analysis when the study was designed. It is obviously best to design the data collection for use with a particular method of statistical analysis. So it is with factor analysis, too. The variables should be selected with some particular theory or conceptual framework in mind, the data should be collected in such a way that appropriate

Requests for reprints should be sent to Andrew L. Comrey, Department of Psychology, University of California, Los Angeles, California 90024.

correlational methods can be used, and a suitable sample of individuals should be selected that will allow for appropriate application of the chosen method of analysis. Unfortunately, all too often, factor analysis is thought of only after the data are already in. At this point, it may be next to impossible to carry out a good factor analytic investigation.

The moral of the story, therefore, is to plan ahead when using factor analysis. The investigator should start with a first-stage conception of what the factor structure should be like in the domain he is studying (Guilford & Hoepfner, 1971; Thurstone, 1947). This first-stage conception may be generated by theory, intuition, past results, or a combination of these elements. The fact that an investigator should start with a tentative conception of what the ultimate factor structure is apt to be does not mean, of course, that he should try to force the results to confirm that original conception. He should at all times be alert to evidence that the data poorly confirm that conception and take steps to revise it accordingly.

Variables should be selected for the analysis that will provide good representative measures for each of the expected factors. Insofar as possible, it is preferable to have variables that will measure one and only one of the expected factors in any substantial way. Complex variables that measure several different factors will be of little value in locating the proper factor structure in the analysis. After the variables have been selected, it is important to develop good measures of these variables that will produce reliable scores with adequate continuous distributions in the sample of individuals to be studied. Without good continuous measures, the correlation matrix to be factor analyzed will be a poor one. Finally, the sample must be carefully selected to be representative of the kind of population to which the investigator wishes to generalize his results. It also must be large enough to give stable correlation coefficients. If possible, there should be at least five times as many variables as the number of expected factors, and there should be a least 200 subjects. I have found continued reduction in the perturbation in factor analytic results up to samples of 2,000 cases before the factor struc-

ture stabilized. Factor analytic results based on small samples can be considered at best as crude hypotheses to be tested in further investigations.

Beyond planning ahead in designing a factor analysis, it is important to understand that the best use of factor analysis takes place in a programmatic series of investigations in which the researcher is constantly refining his conception of the factor structure and the variables that represent the factors, improving the extent to which the variables measure the factors, and testing hypotheses in subsequent studies generated by results in prior investigations. A hypothetico-deductive approach is recommended. The initial investigation gives a first approximation to what the factor structure should be. On the basis of this conception, modifications in existing variables are made, new variables are added, and so on, and predictions are made about what will happen in the next study as a result of these changes. The next study verifies or fails to verify these predictions, thereby generating further suggested changes. In this manner, the researcher gradually clarifies what the factors are in a given domain and what variables measure them in the most effective way. His results are stabilized over a series of investigations using different samples.

This is not to say that a one-shot, after-the-fact treatment of existing data using factor analysis in a hunting expedition is always useless. Such results are useful, however, mainly as hypotheses to be followed up by more adequate programmatic research. In and of themselves, they have little value from a scientific standpoint unless they can be verified in other investigations. It should also be emphasized that good programmatic factor analytic research usually will not be sufficient in and of itself to establish the ultimate utility of the identifying factors. Internal analysis of correlational data usually must be supplemented by experimental and other kinds of non-factor-analytic investigations to test hypotheses related to the factors before their scientific value can be clearly demonstrated.

Selection of the Sample

The particular sample of individuals studied in a factor analytic investigation can have a

profound effect on the outcome of the analysis. If we give ability tests to a sample of children ranging in age from 8 to 14, for example, a huge general factor may be generated, since most of the 14-year-olds will probably do better than most of the 8-year-olds on all the tests. The correlations between all tests will be very high, and a single factor will account for most of the variance. An unwary investigator might call this a general intelligence factor, but in reality it is only a maturation factor. If all the tests were to correlate substantially within a sample of children of the approximate *same age*, on the other hand, this would be indicative of a general intelligence factor running through these cognitive measures. If, however, students from a very selective educational institution of higher learning were given tests of a similar kind, the general factor might disappear entirely because variance would be so restricted on general intelligence. Since all of the students would be very bright in such a sample, the general factor would be lost, leaving only group factors.

It is very important to avoid idiosyncratic collections of subjects that can have gross effects on the correlation matrix. For example, if two subjects who have much higher (or lower) scores than all of the other subjects on several of the variables are in the same sample, this will greatly elongate the scatter plot ellipses, generating very high correlations for those pairs of variables and injecting a great deal of spurious common factor variance into the matrix. The investigator cannot necessarily anticipate all of the special circumstances that might give him spurious results. He can be suspicious, however, of any unexpectedly high correlations, or low ones for that matter, and investigate thoroughly to determine if there are special problems with his data leading to distortions in the correlation coefficients.

Selection of the Variables

It has already been indicated that a preliminary conception of the probable factor structure in the domain under investigation should be formulated, and then variables should be selected for the analysis that will

provide good definitions of these putative factors. Ideally, a variable for a factor analytic study should measure one and only one factor in the domain to any substantial degree; that is, it should be a pure-factor measure. This is not usually possible for all of the variables in the matrix, since it is difficult to develop pure-factor measures. In particular, it is not likely to be the case when the investigator is bound to use existing data that were not collected with a factor analysis in mind. When the analysis is planned in advance, the investigator can and should carefully plan which variables he wishes to include, selecting approximately an equal number of variables for each hypothesized factor and trying to make each variable as pure a measure of one factor as he can. He should seek particularly to avoid the situation in which a substantial percentage of the included variables are complex measures, each with substantial expected contributions to two or more factors. It is very difficult, if not impossible, to attain an appropriate rotational solution with any known analytic method when a majority of the variables in the analysis are complex.

Beyond getting the most appropriate collection of variables in the matrix, it is very important to develop good, reliable measures of those variables. All too often, there is an attempt made to factor analyze poor data variables, such as two-choice questionnaire items with poor splits in the proportion of yes and no responses. Proper variables for factor analytic purposes should be relatively continuous; that is, they should have many possible categories of response and reasonably normal distributions. Also, the regressions for every pair of variables should be linear and reasonably free of gross departures from the normal expected elliptical pattern of data points in the scatter diagram. When two-choice response variables are used, it is not even possible to investigate linearity, and furthermore severe distortion can be introduced into the correlation matrix with a consequent dramatic effect on the factor analytic solution. If one variable is represented by a measure that splits 50-50 while another is represented by a measure that splits 95-5, the maximum possible correlation between the two variables is limited to an absolute value of

approximately .23. With more appropriate measures on continuous scales, these two variables might correlate much higher. The form of measurement of these two variables imposes an artificial limit on the size of the correlation that could introduce a serious distortion in the obtained factor structure.

Even more dramatic distortions may occur, however, because grossly inflated correlation values also can be introduced by poor measures of the variables used. Again using two-choice variables as an illustration, suppose that in a sample of 200 cases, 199 subjects answered no to two items and one person answered yes to both. In this case, the correlation between the two variables would be 1.00 because of 1 deviant individual out of 200. If that person had said no to both items, the correlation would have been .00 instead of 1.00. Thus, with two-choice data, the correlations can be artificially limited in size or they can be grossly inflated, depending on the situation encountered, as compared with what probably would have occurred with continuously measured, normally distributed variables. The investigator, therefore, should seek to use variables for analysis that are continuous rather than dichotomous. I have personally given up using two-choice items wherever possible, preferring to work with seven-choice items if they are to be factor analyzed. This substantially reduces the possibility of gross distortions in the correlations like those indicated above. Beyond this, in factor analytic studies with taxonomic objectives, I prefer to use variables that are total scores summed over several homogeneous items that have been shown to be measuring the same variable. These total scores have many more possible data points than the seven associated with a single item. Such scores are more continuous, tend to be more normally distributed, and are almost always more reliable than individual item scores. Factor analytic results for such improved measures are much less subject to distortion than those based on dichotomous measures from single items.

Attempts have been made to correct for errors in two-choice variable intercorrelations by using such devices as phi over phi max and tetrachoric correlations, but these methods even theoretically only deal with the

problem of artificial limits on the size of the correlations. These methods do not even address the problem of spuriously high correlations. Comparisons of the results of factor analyses of the same data using phi coefficients (Pearson correlations with 1 and 0 values assigned to the two dichotomous categories), phi over phi max, and tetrachoric correlations show that the latter two coefficients can grossly inflate the amount of common factor variance in the matrix in certain situations (Comrey & Levonian, 1958). The sad fact is that nothing can be done to make good data out of poor data. Wherever possible, the investigator should make plans in advance to obtain continuously measured scores for the variables that he wishes to factor analyze. If dichotomous measures must be used, try to get them as close to 50-50 splits as possible for all measures and then interpret the results very cautiously. In particular, look for any large correlations that might have been artificially elevated by idiosyncratic response distributions in the measures correlated.

Extraction of Factors

What to put in the diagonal cells as initial communalities and how many factors to extract are the two decisions that create the most difficulty for typical research users of factor analysis. Many popular computer programs offer the option of inserting 1.0 for each communality (diagonal) cell of the correlation matrix followed by the extraction of all factors by the principal factor method that have eigenvalues greater than or equal to 1.0 (called the "eigenvalue-one criterion" by Rummel, 1970, p. 363). That is, the sum of the squares of all factor loadings for the extracted factor is 1.0 or greater. One justification for this procedure is that each variable included in the matrix adds 1.0 to the total communality in the entire matrix, hence any retained extracted factor should contribute at least as much as the effect of adding one variable to the matrix. This option has become very popular, probably because it is readily available in existing computer programs that are widely used, and because it requires no judgment on the part of the investigator.

There are problems with this approach, however, that can lead the unwary investigator astray (Comrey, 1973; Guertin & Bailey, 1970; Lee & Comrey, Note 1; Comrey & Lee, Note 2). First of all, each variable has 1.0 as its potential communality after all possible factors have been extracted. This is true whether the variable has much or little actual common factor variance. By the time all factors have been extracted with eigenvalues greater than 1.0, it not infrequently happens that an amount of common factor variance has been extracted for certain variables that is inappropriately high given the size of the correlations with the other variables in the matrix. This is particularly apt to occur when items of a two-choice nature are being correlated. These variables are notoriously unreliable and often have low correlations with each other. The factor results based on analysis of such matrices with the eigenvalue-one procedure described above, however, often show very high communalities and large factor loadings for these dichotomous variables. When this spurious extra common factor variance is introduced into the matrix, the number of factors that seem worthy of retention even after rotation, is often too high, further distorting the solution. It is my conclusion that the eigenvalue-one procedure should not be used unless the variables in the matrix are all good, reliable, continuous measures that correlate substantially with each other, leading to rather high true communalities for all variables. In this case, the distortion introduced by the eigenvalue-one procedure is less likely to represent a problem.

The distortion introduced by this method can be mitigated by iterating the communalities, that is, after extracting factors with eigenvalues greater than or equal to 1.0, reinsert the accumulated communalities for that number of factors, reextract the same number of factors, and repeat (Lee & Comrey, Note 1; Comrey & Lee, Note 2). Unfortunately, the values to which the communalities converge, if they do, will not necessarily and often will not be, the same values to which they would converge starting from different initial communality estimates. If unities are used in the diagonals to begin with for a common factor analysis, however, I believe that it is better to

iterate the communalities to stability, since this will typically introduce less distortion into the solution than failing to iterate.

It cannot be assumed, however, that the correct number of factors is obtained by stopping extraction when the eigenvalues are less than 1.0. This procedure would seldom give too few factors, although it often gives too many. The investigator should look at a number of criteria for terminating extraction, including examining rotated solutions with different numbers of factors, before leaping to the conclusion that he has the right number. There is no definitive solution to the problem of determining the correct number of factors. It is up to the investigator to use all of the information that he can get to reach a conclusion about this and to try to justify his conclusion within several converging lines of evidence instead of relying on a universal rule of thumb.

A second common factor extraction strategy is to use squared multiple correlations in the diagonal cells and extract all factors by the principal factor method having eigenvalues greater than zero. This procedure can be followed by iteration of the communalities until they stabilize. The stabilized communalities generally will not diverge from the squared multiple correlations by this procedure as much as the stabilized communalities will differ from 1.0 in the previously discussed procedure. For this reason, iteration of the communalities is less necessary in this latter procedure, and in fact, it may not give a better solution than the noniterated communalities. Iteration of the communalities tends to increase the size of these values, on the average, capitalizing on chance error to raise the proportion of variance extracted. I have found cases in which it appears that the communalities have been raised to unrealistically high values by this iterative process (Comrey, 1973). It is not unreasonable, however, to iterate the communalities and examine the results in comparison with the uniterated results to see which solution can be justified as being the more reasonable. The number of factors must be kept constant during the iteration process, however, or the number of factors will gravitate toward the number of variables and the communalities will gravitate toward 1.0

instead of stabilizing at smaller values. Using squared multiple correlations and eigenvalues greater than zero as a criterion will typically extract substantially more factors than the eigenvalue-one procedure. Care must be taken to eliminate superfluous factors during the rotation process to avoid distorting the final solution.

A third alternative factor extraction strategy is to obtain a solution that does not require communality estimates. Comrey (1962) and Comrey and Ahumada (1964) have proposed a stepwise minimum residual solution that operates only on the off-diagonal elements of the correlation matrix. Communalities are derived as a result of the factor extraction process rather than being estimated ahead of time and then influencing the calculation of factor loadings. The Comrey and Ahumada procedure has a built-in criterion for terminating extraction when iterated factor vectors converge on vectors of opposite sign in the extraction process. This corresponds roughly to the point in the principal factor method of terminating factor extraction where eigenvalues become negative when squared multiple correlations have been used as initial communality estimates. The communalities that result from this minimum residual procedure can be used as the point of departure for an iteration of communalities by the principal factor method. This may or may not prove to be defensible, depending on the degree of elevation of the communalities by the iterative process. Iterating the communalities, starting with the minimum residual solution communalities as initial estimates, appears to give about the same results in most cases that would be obtained for that number of factors using the Harman and Jones (1966) minres method. Minres is also a type of minimum residual solution that reduces the off-diagonal residuals to a minimum, but it does so after a specified number of factors, not one at a time as in the Comrey (1962) method. With the minres method, the number of factors desired must be specified in advance.

Thus, extracting factors by the principal factor method with squared multiple correlations in the diagonal and iterating until communalities stabilize, extracting factors by the

Harman and Jones (1966) minres method, or extracting factors by the Comrey method (1962) and iterating the communalities all give similar results if the number of factors is held constant. Extracting the same number of factors with the communalities set equal to 1.0 can give very different results that are difficult to defend. If the eigenvalue-one results are iterated until the communalities stabilize, the amount of distortion will typically be lessened although not necessarily eliminated entirely. Determining the number of factors to extract is still a major problem with all of these methods. Ordinarily, it is necessary to consider a variety of criteria and even rotate various numbers of factors before coming to a final conclusion about what is the appropriate number of factors. Just accepting the number of factors suggested by a particular termination criterion can lead to gross distortions in the final results.

Many other procedures for extracting factors are available, each with its own characteristics and potential abuses.¹

Rotation of Factors

The unrotated factor matrix is computed in most methods such that the product of this matrix by its transpose will approximate the original correlation matrix within a certain margin of error. If a given factor matrix will do this, so will an orthogonal transform of that matrix, that is, a matrix that can be reached by orthogonal rotations of the unrotated factors. The axes do not even have to be rotated orthogonally, but then the angles formed by the axes must be considered in reproducing the correlation matrix. The various methods of factor extraction obtain the unrotated matrix by satisfying certain mathematical criteria, such as extracting the maximum amount of variance in one factor or minimizing the sum of squares of residuals of off-diagonal elements, and so on. In almost all cases, these mathematical criteria locate the factor axes in positions that have nothing to do with psychological meaningful positions for the axes. In these cases it is necessary,

¹ For further information, consult the factor analysis textbooks mentioned later in the article.

therefore, to carry out rotations of the unrotated factor axes, either orthogonal or oblique, before the factor axes are appropriately located to define psychologically meaningful constructs. Since the number of different positions for these axes is virtually unlimited, there is no unique solution to the rotation problem in factor analysis. Many different criteria have been developed to help the investigator locate the "correct" factor axis positions, but none of these criteria has achieved general acceptance. The basic reason for this is that no one criterion of rotation *can* fit all situations, since the proper location of axes in a given matrix is often quite specific to that particular problem. A rotational solution is nothing more than the investigator's interpretation of the data. It may or may not be a good one.

Thurstone (1947) popularized the criterion of simple structure for rotation of factor axes, and many of the modern computer procedures for rotation of axes are merely attempts to approximate simple structure through the application of specially designed mathematical algorithms to the data. Simple structure usually will work reasonably well, as will computerized approximations to it, if the study is properly designed for the use of this criterion. Thus, if there are several well-defined factors, each measured by several pure-factor measures (one-factor variables) that are normally distributed and reliably assessed, and if there are few if any complex factor measures, simple structure will ordinarily work very well indeed. Unfortunately, in practice relatively few factor analytic studies carried out actually fit this model. In most actual analyses, the factors are not so well charted in advance, each factor does not have several pure-factor measures to define it, and there are usually many complex variables. The number of variables per factor may vary considerably, and often there are too few well-defined factors in the matrix.

When the factor matrix being analyzed falls into this latter category, simple structure is not apt to be very successful in pointing the way to the best solution of the rotation problem, and of course no computer algorithm designed to approximate simple structure will either. A problem such as this requires all of

the skill and experience that the most knowledgeable factor analyst can muster to have any hope of locating the best solution. Since such idiosyncratic collections of variables do not fit the simple structure paradigm, to locate the best factor structure for the data it is often necessary for the investigator to fix the rotated factor positions in accordance with knowledge that he may have about the variables. Thus, he might know that one variable has been found in several past analyses to be a relatively pure measure of a certain factor. In his rotations, he would deliberately place one of the factors coincident with this variable vector and attempt to rotate it in such a way that variance for this variable would not appear in any substantial way on other factors. He might know that two other highly correlated variables were complex composites of factors X and Y. Not having pure measures of X and Y, positioning of the axes would be difficult, but at least the factors could be positioned in such a way that a factor is not run through these two complex variables as an analytic rotation program might. In short, a resort to the old "hand-rotation" procedures using all of the knowledge that one can muster about the variables may be necessary to have any chance of getting a reasonable rotated factor solution (Comrey, 1973). This skill has almost become a lost art with the advent of the high-speed computer. These hand rotations can be performed without undue labor, however, using the computer to do the hard work, so there is no need to give up this valuable option just because it used to be so laborious.

I have developed a computer program (Comrey, 1973), for example, that first plots the factors. These plots are inspected visually to find the desired rotation, orthogonal or oblique, and then the instructions are given to the computer to carry out the rotations and make new plots. The investigator only has to inspect the plots and determine which factors he wishes to rotate by which angles of rotation. This is more work than applying a computer-programmed analytic method, of course, and it requires the application of some judgment, but it is not laborious, and it is a skill that can be mastered with some practice.

Critics of factor analysis will be quick to

point out that such a procedure is subjective and hence suspect. This is no doubt true. The investigator is charged with finding the best possible interpretation of the data that he can and then justifying his interpretation with evidence that goes beyond the factor analytic techniques used to obtain that solution. Application of a mathematical algorithm is more objective to be sure, but it does not alter the onus placed on the investigator to justify by other lines of evidence that he has, in fact, found the best structure for the data. If the mathematical algorithm applied happens to be inappropriate for the data and happens to give nonsense for results, it is small comfort to the investigator to know that the process was untainted by human judgment. It is the fond wish of naive investigators that factor analysis should be an objective machine for grinding out truth untouched by human brains. Such fairy tale devices do not exist in the behavioral sciences, nor indeed in the physical sciences. The reality is that factor analysis is somewhat like a microscope. It can help an intelligent and well-trained investigator see things that he could not see with the naked eye, and perhaps derive scientific conclusions of value as a result. The untrained individual looking through a microscope can see many things, but he may not be able to interpret what he sees correctly. The untrained factor analyst also can get a final rotated factor matrix from the computer center, but he may not be in a very good position to draw appropriate conclusions from what he sees. Just as we would not throw out or blame the microscope because an untrained individual uses it improperly, we should not blame the tool itself when factor analysis is applied improperly.

Orthogonal versus oblique axes. The decision to use orthogonal or oblique rotated axes should be a conscious one by the investigator for good cause, not based on just what happens to be available at the computer center. Many investigators give no thought to determining whether they should have an oblique or orthogonal solution and if oblique, how oblique. Conformity to simple structure (more variables with low loadings) can be improved more and more as axes are allowed to become more and more oblique. Some ana-

lytic procedures require the specification of a constant that will put more or less of a brake on this tendency toward greater obliqueness. Others require the specification of a maximum permissible correlation between oblique factors. To avoid inappropriate degrees of obliqueness in the solution, whatever rotational method is used, the investigator should have some idea of what correlations between his factors are reasonable. The actual angles between factors in his oblique solution should be examined with this in mind. If the oblique solution rendered by an analytic computer method yields factor correlations that cannot be justified as reasonable, the investigator should be prepared to modify the solution.

Rotating different numbers of factors. The number of factors rotated in a solution can have a profound effect on the results, especially if a computerized mathematical algorithm is applied. A common error is to extract too many factors and then rotate them all by varimax or some other procedure that seeks something approximating a simple structure type of solution. The varimax algorithm will build up minor factors where possible at the expense of major factors. Especially where the eigenvalue-one procedure is used and too many factors are rotated, the varimax method may produce one or more factors with high loadings for one and only one variable. With a different kind of rotation procedure, this variable might have an important loading on one of the major factors but not when the varimax procedure was used with too many factors being rotated. The presence of a rotated factor with one large loading, say .6 or more, and no other loading above .35, is almost a certain sign of a distorted factor solution when common factor variance is being analyzed.

Skillful hand rotations can avoid the problems associated with rotating too many factors. The extra small factors are simply not allowed to distort the proper appearance of the major factors. It is analytic rotation of these extra factors by mathematical algorithms that is apt to cause a problem. One way of dealing with this problem is to rotate several numbers of factors in the region of uncertainty by the mathematical algorithm

in question and then compare the solutions to see which can be justified best as a proper interpretation of the data. I have developed a procedure called the "tandem criteria" method of rotation (Comrey, 1967) that starts out by rotating more than enough factors. Variance is shunted from the major factors toward the smaller factors, but this process is restrained by the correlations that exist among the variables. Correlated variables are retained on the same factors. After this process, factors that are still too small are dropped. Retained factors are rotated by a different algorithm to provide a more simple structure type of solution. It is often necessary to rotate two or three different numbers of factors in the region of uncertainty before a final decision is made about the "correct" number of factors.

Rotating too few factors, on the other hand, forces the amalgamation of factors into complex composites, obscuring the most useful factor structure. Variance is ordinarily extracted in successively smaller amounts from the first to the last factor, but parts of the variance from several factors appear as each factor is extracted. If only 10 factors are rotated when there are really 12 factors there, most of the variance for the two extra factors will be included in the first 10 unrotated factors. This variance must go somewhere, so it is superimposed on the first 10 factors, distorting their true appearance. Nothing can be done to correct the effects of rotating too few factors, except rerotating with more factors. It is better, therefore, to err on the side of taking out too many factors rather than too few.

One of the main points of this discussion is to suggest that it is often impossible to determine the correct number of factors just by considering the unrotated factors alone. Rotational solutions can often provide important information that will be useful in making this decision. In most empirical analyses, particularly with small samples, there is a good deal of error in the correlations that tends to add surplus variance to the solution. Guertin and Bailey (1970) have pointed out that the larger variability of chance correlations around zero for small samples compared with large samples produces greater commu-

nalities, and in many cases this can mean extra factors. Trying to decide just where the wheat ends and the chaff begins is not easy. The decision should take advantage of every scrap of evidence available to the investigator, including the benefit of comparing solutions with different numbers of rotated factors. This increases the dangers from subjectivity, of course, but for a cautious, scientifically oriented investigator, the benefits outweigh the disadvantages in my opinion.

Confusing factor levels. A persistent problem that plagues many factor analytic investigations is the inadvertent production of low-level factors in the hierarchy of factor generality. A factor of this kind is indicated by the presence on the factor of two or more variables with high loadings that are very similar in character. Thus, systolic and diastolic blood pressure both included in the same analysis will produce such a low-level factor, or at least distort the nature of some factor, depending on the other variables included in the analysis. This is due to the fact that these two variables are highly correlated, measuring very similar phenomena. As another example, consider these two test items (a) I feel blue, and (b) I am depressed. Correlating these items with others and factor analyzing them will produce a low-level factor, with these two items showing high loadings. These two variables are merely alternate forms of the same thing. Any factor analytic solution should be inspected for the presence of such low-level factors. They cannot represent constructs of general interest and scientific utility per se, because these factors can be produced virtually at will. It is only necessary to add another variable to the matrix that is very similar to an existing one, and then a factor will be produced defined by these two variables.

In my opinion, factors are more apt to be useful if they stand at the next level up in the hierarchy of factor generality. The variables with high loadings on the factor should not contain two or more variables that could represent virtual alternate forms of the same thing. Each variable with a high loading on the factor should have a logically distinct and separate identity, measuring something that is not the same as what is being measured by

any other variable on the factor. The high-loading variables defining the factor should be correlated but not alternate forms of each other. In my experience, it is not possible to keep adding factors of this kind in a given domain. One very quickly comes to the end of the line. There is, on the other hand, no end to the number of low-level factors that can be generated. This fact alone favors second-level factors versus first-level factors. At the other extreme, also to be avoided, are factors that are too high in the hierarchy. Factors produced by correlations among highly complex variables, such as several tests of intelligence, are too broad to have much utility. When several second-level factors are found to be persistently intercorrelated, however, a useful third-level construct may be indicated. This is the case with general intelligence, a third-level construct that is generated by the intercorrelations that exist among lower level factors of cognitive performance. Identification of such useful higher order constructs takes place best through the identification of second-level factors that are correlated, not by a direct assault using complex variables in the factor analysis.

Locating a general factor. Occasions arise when an investigator hopes to demonstrate the existence of a general factor running through the variables that he is investigating. One of the common errors in this instance is to take the first extracted factor and treat it as though it were a general factor. This procedure is unsatisfactory, because the first factor vector can easily contain major loadings for entirely unrelated variables. If the factor is to be a general factor, the variables relating to it must be correlated with each other, except in rare special cases involving certain kinds of complex variables. The investigator should make sure that the variables loading on his putative general factor are indeed correlated to a satisfactory degree before he leaps to the conclusion that he does in fact have a general factor. On the other hand, if rotations of several factors are carried out by varimax, for example, or any other method seeking simple structure, the tendency will be to disperse the variance as much as possible, thereby reducing the prominence of any general factor that might be present. It is

desirable to use a method that will permit a general factor to appear if in fact the data warrant it without forcing it on the data. The Criterion I rotational method of the tandem criteria (Comrey, 1967, 1973) is one such method.

Interpretation of Results

The usual procedure in interpreting factor results is to inspect the variables that have high loadings on the factor, look for what they have in common, and then name the factor in accordance with the common elements. Up to this point, the activity is little more than factor naming, and if nothing beyond this is done, the value of the analysis may be rather limited. When a name is given to a factor, a hypothesis has been formulated. Untested hypotheses usually have limited value until something has been done to test them. Ideally, as has been mentioned earlier, the investigator will make plans to carry out additional analyses in which he adds new variables that should have major loadings on certain specified factors and low loadings on other factors if his hypotheses are correct. He will perhaps revise other variables in ways that predict certain outcomes if his factor interpretations are correct. These predictions will be tested by further investigations. Experiments may be carried out, with predictions being made as to the outcome, in which attempts are made to alter scores on one factor but not another. Results of the experiment will confirm or disconfirm the hypothesized factor meanings. In other cases, predictions may be made about how scores for a certain factor will correlate with other variables outside the matrix. When the investigator carries out one analysis and does not follow up after the factor naming to test the accuracy of his factor interpretations, he must recognize that much remains to be done before his interpretation can be regarded as proved.

Writing Up Results

One of the most commonly violated rules of scientific reporting in factor analytic studies is that the writer fails to give enough information to permit another investigator to

repeat his work. There should be sufficient information given about the variables studied so that someone else can also construct or obtain comparable measures if he so desires. The correlational matrix itself should be made available to the reader so that he can evaluate the adequacy of the solution reported or carry out his own preferred type of analysis on the data. Future developments may also provide more powerful methods of analysis that can be applied to the same matrix, provided that the matrix is made available. Many factor analyses appear in manuscripts submitted for publication that fail to contain this information, making it exceedingly difficult to evaluate the work. True, these matrices often cannot be published in the article itself, but at least they can be made available through auxiliary publication outlets and certainly should be available to journal editors.

Beyond this minimum, the investigator should preferably do much more and often fails to do so. He should report enough about his sample so that its impact on the results can be assessed. He should tell what method of carrying out the factor extractions he used, what he did about the communalities and why, and how he determined the number of factors extracted and rotated. He should also make available the unrotated factor matrix, using auxiliary publication outlets.

The method of carrying out the factor rotations should be explained and the reasons given why this is an appropriate method for the data. Some information should be given about why an orthogonal or an oblique solution was chosen. Discussion of the rotational solution with respect to the problems of determining the correct number of factors should show that the investigator at least considered alternate possibilities. The rotated factor matrix should also be made available. If there is an oblique solution, the structure matrix, the pattern matrix, and also the matrix of correlations between factors should be made available. Again, auxiliary publication outlets should be used for these matrices, but the journal editors should have access to these materials at the time of manuscript review.

Sources of Help

In an article such as this, it is possible to point out briefly some common trouble spots in factor analytic investigation and reporting, but it is obviously not possible to say all that needs to be said nor to give adequate advice on how to avoid all possible pitfalls. Several good textbooks are available to help the investigator, each with its own strengths, weaknesses, and special emphasis. A few of the more recent books are mentioned below with brief annotations:

Comrey (1973) assumes a knowledge of beginning statistics and high school algebra and gives elementary treatment of the mathematical background of factor analysis, but the major emphasis is on practical applications. Computer programs are made available.

Gorsuch (1974) contains a balanced approach between theory and applications with a level of difficulty greater than Guertin and Bailey (1970) but less than Harman (1976), Horst (1965), and Mulaik (1972). It is a good introduction to the more difficult books.

Guertin and Bailey (1970) is essentially nonmathematical in its presentation and hence tends to be a "how to" cookbook; despite its weakness on theory, the book contains much useful information and results of many experiments with different factor methods.

Harman (1976), published posthumously, is the latest edition of Harman's widely used text and is an excellent source for the mathematical methods and theory of factor analysis. It is difficult reading for the person who is poorly trained in mathematics, however, and it does not emphasize the practical problems in applying factor analysis.

Horst (1965) is an encyclopedia of information about theory, methods, and techniques in factor analysis. Unfortunately, Horst's notation makes the book very difficult to read for most behavioral scientists.

Mulaik (1972) is a thorough mathematical treatment of factor analytic techniques. The major emphasis is on mathematical theory rather than applications. Unless the reader is fairly sophisticated mathematically, he should probably start with some other book.

Rummel (1970) is a good introduction. It emphasizes applications more than theory and

is similar to the Comrey (1973) and Gorsuch (1974) texts in level of difficulty and mathematical preparation required.

Reference Notes

1. Lee, H. B., & Comrey, A. L. *Distortions in a commonly used factor analytic solution*. Paper presented at the annual meeting of the American Psychological Association, San Francisco, August-September 1977.
2. Comrey, A. L., & Lee, H. B. *Distortions in some commonly used factor analytic solutions*. Unpublished manuscript, University of California, Los Angeles, 1977.

References

- Comrey, A. L. The minimum residual method of factor analysis. *Psychological Reports*, 1962, 11, 15-18.
- Comrey, A. L. Tandem criteria for analytic rotation in factor analysis. *Psychometrika*, 1967, 32, 143-154.
- Comrey, A. L. *A first course in factor analysis*. New York: Academic Press, 1973.

- Comrey, A. L., & Abumada, A. An improved procedure and program for minimum residual factor analysis. *Psychological Reports*, 1964, 15, 91-96.
- Comrey, A. L., & Levonian, E. A comparison of three point coefficients in factor analyses of MMPI items. *Educational and Psychological Measurement*, 1958, 18, 739-755.
- Gorsuch, R. L. *Factor analysis*. Philadelphia, Pa.: Saunders, 1974.
- Guertin, W. H., & Bailey, J. P. *Introduction to modern factor analysis*. Ann Arbor, Mich.: Edwards, 1970.
- Guilford, J. P., & Hoepfner, R. *The analysis of intelligence*. New York: McGraw-Hill, 1971.
- Harman, H. H. *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press, 1976.
- Harman, H. H., & Jones, W. H. Factor analysis by minimizing residuals (minres). *Psychometrika*, 1966, 31, 351-368.
- Horst, P. *Factor analysis of data matrices*. New York: Holt, Rinehart & Winston, 1965.
- Mulaik, S. A. *The foundations of factor analysis*. New York: McGraw-Hill, 1972.
- Rummel, R. J. *Applied factor analysis*. Evanston, Ill.: Northwestern University Press, 1970.
- Thurstone, L. L. *Multiple factor analysis*. Chicago: University of Chicago Press, 1947.

Received August 22, 1977 ■

Experimental Methods and Outcome Evaluation

Michael J. Mahoney
Pennsylvania State University

Within the constraints imposed by philosophy of science and the limitations of scientific methodology, experiments are basically attempts to identify causes and to evaluate their effects. All experiments involve a comparison of dependent variable values in the presence of different values of another variable. This comparison is usually accomplished by one of two basic experimental strategies: (a) within-subject designs and (b) between-subject designs. The 12 most common designs are outlined and discussed in the present article. Much of the discussion focuses on the importance of internal, external, and theoretical validity in experimentation and data interpretation. To illustrate the most common threats to these three forms of validity, various components in a hypothetical experimental manuscript are discussed. This discussion is followed by an acknowledgment of the continuum of fallibility along which all experiments fall. It is argued that in the final analysis, our goals should be to strive toward conducting the least fallible inquiries, to cautiously interpret our experiments in accord with their logical warrant, and to guard against the paralysis of complacency regarding the adequacy of current research methods.

The methodologist is often stereotyped as a person who is quite willing to offer criticism but who is generally less adventurous when it comes to the actual execution of research. Although armchair quarterbacking is a common diversion among scientists, the bulk of our strategic insights seems to follow rather than precede the busy work of experimentation. Besides illustrating the old maxim that hindsight is more plentiful than foresight, our post hoc criticisms also seem to exhibit another common theme—namely, that we tend to be more critical of other people's research than of our own. It is, therefore, with some reservations that I comment on experimental problems in clinical psychology. I do not presume that my own research has been methodologically flawless, nor do I harbor many illusions about the limitations of science. The perfect experiment has yet to be designed and is, in some sense, inconceivable (Weimer, 1977). Even if it were conceivable, however,

it is a safe bet that it would be impossible to execute. Among other things, the human element in science makes it an inevitably fallible endeavor (Mahoney, 1976; Mitroff, 1974). Let us therefore dismiss the notion of an ideal experiment and instead devote our attention to the continuum of fallible effort along which all experiments must fall. In this regard, our past mistakes and costly hind-sights may well serve to refine our continuing attempts to improve the field of clinical psychology.

As clinical researchers, it is our goal to harness the powers of science in a manner that will help us refine human services—particularly psychotherapy and counseling. This might sound like a rather straightforward undertaking in that it would seem to require only a knowledge of scientific methods and the opportunity to apply them to problems of human distress. As it turns out, however, few undertakings are quite as ambitious as the scientific study of therapeutic interaction. This may be due in part to our poor understanding of the nature of science (Lakatos & Musgrave, 1970; Weimer, 1977). At this point in time, for example, no acceptable cri-

Requests for reprints should be sent to Michael J. Mahoney, Department of Psychology, 417 Bruce V. Moore Building, Pennsylvania State University, University Park, Pennsylvania 16802.

erion has been established for the demarcation of science from nonscience. We must therefore proceed, if at all, with a humility about our understanding of what it is that constitutes "good scientific research" and show considerable tolerance for statements that convey relativity and tentativeness. Today's science may well be tomorrow's alchemy, and it is imperative that this contextual feature be appreciated. With these apologetics in mind, we can now proceed to a discussion of contemporary issues that bear on the evaluation of research.

Fundamentals of Experimental Method

In at least one sense, science is a search for order—an attempt to discern, describe, and apply systematic covariations between events. Order is a prerequisite for prediction, and accurate prediction is a first step toward control. This is why the assumption of determinism cuts so deeply into the core of scientific methodology (Hook, 1958). Although systematic covariation is a prerequisite for accurate prediction, however, it is hardly sufficient for the inference of a cause-effect relationship. This, of course, relates to the oft-quoted difference between correlation and causation, or what David Hume regarded as the distinction between *sequence* and *consequence*. Because covariation, no matter how systematic, can never demonstrate causation, correlational inquiries (or "associative" studies) are often considered nonexperimental in nature. This is true, of course, if one defines an experiment as requiring an active manipulation of one or more variables. It should be kept in mind, however, that associative inquiries are hardly uninformative. Although they can never conclusively demonstrate causality, this is also true of the best conceived experiment. The difference between the two is not as dichotomous as some might expect, with both types of inquiry varying in the relative strength of the conclusions they warrant. Moreover, correlational studies *can* disconfirm some hypothesized relationships, and in this sense they are capable of corroborating a hypothesis. There is a tremendous difference between corroboration and confirmation, however, which we shall explore in a

moment. For the time being it is worth noting that such sciences as astronomy and meteorology are almost exclusively correlational, yet their progress is unmistakable. Having made this brief defense of associative inquiry, I shall hereafter restrict my comments to the topic at hand—namely, problems in *experimental* research as it has been traditionally conceived.

There are, of course, substantial parallels and yet important differences among the concepts of prediction, causation, and explanation. Since the technical aspects of these terms have consumed entire volumes, only the most superficial rendering could be offered here. I shall therefore defer to Weimer's (1977) more extensive discussions, which bear direct and significant implications for the conduct of research. In lieu of a more technical digression, suffice it to say that contemporary scientists often require the following in their evaluation of a causal relationship:

1. relative temporal contiguity (togetherness in time);
2. priority (the cause must precede the effect);
3. noncontradiction (no observed instances of the cause without the effect);
4. factor isolation (the elimination or control of all possible influences other than the one being examined); and
5. replicability (the capacity to replicate the alleged relationship).

The logical validity of these criteria will not be examined here, but this should not be interpreted as complacency regarding our understanding of the concept of causation.

Scientific experiments are basically attempts to identify causes and to evaluate their effects. This is usually accomplished by making systematic changes in one or more factors (the "independent variables") and looking for any covarying changes in other factors (the "dependent variables"). The division of factors into independent and dependent categories is often arbitrary, and in some inquiries auxiliary variables are measured (e.g., to evaluate whether they can help predict the prerequisites or range for a cause-effect influence). Common to all experiments, however, is the act of comparison. *All experiments involve a comparison of dependent variable*

values in the presence of different values of another variable. In extreme form, the manipulated (independent) variable may take on dichotomous values (e.g., present vs. absent). The basic question, of course, is whether different values of the dependent variable(s) are associated with different values of the independent variable(s). This comparison is usually accomplished by one of two basic experimental strategies: (a) within-subject designs and (b) between-subject designs.

Within- and Between-Subjects Designs

As the term implies, within-subject designs are ones in which the comparison is made *within* the same subjects across time. That is, the dependent variable is measured at least after the introduction of the independent variable, and perhaps before or during experimental intervention. This may be done with a small number of individuals, in which case it is called "single subject" or " $N = 1$ " research. The merits and shortcomings of single-subject research have received considerable attention (cf. Greenwald, 1976; Hersen & Barlow, 1976; Kazdin, 1973; Sidman, 1960; Thoresen, in press). Within-subject comparisons can also be performed en masse on a single group of individuals, however, so that the within-subject design is not necessarily restricted to studies involving very few subjects.

Between-subjects designs also attempt to compare values of the dependent variable(s) in the presence of different values of the independent variable(s). In these experiments, however, the primary strategy is to look at differences *between* persons who, for example, have and have not been exposed to the independent variable(s). Historically, most between-subjects designs have involved comparisons between groups rather than individual subjects, and thus the between-subjects design is often equated with "group research." It should be kept in mind, however, that both involve the aforementioned comparison. There are many varieties of both within-subject and between-subjects designs. Many of the former are outlined in Hersen and Barlow (1976), and between-subjects designs are extensively described in Campbell and Stanley (1963) and Paul (1969). The 12 most com-

mon designs will be outlined later, but we should first discuss three types of experimental validity.

Validity: Internal, External, and Theoretical

Few terms in scientific research have been applied as widely as the word *validity*, which, to make matters worse, takes on different meanings in its various uses. Without listing its numerous other uses and varieties, the term *validity* will here refer to "logical warrant" or the extent to which a statement or procedure is logically consistent with the experimental intent. Borrowing from Campbell and Stanley (1963), *internal validity* refers to the extent to which a given set of procedures allows one to draw valid conclusions about what actually happened in an experiment. When an experiment exhibits maximal internal validity, one has relatively strong logical warrant for deciding (a) whether an "effect" occurred and (b) whether that effect can be attributed to the independent variable(s) in question. *External validity*, on the other hand, refers to the extent to which a set of experimental procedures allows one to draw valid generalizations to other subjects or situations. A cause-effect relationship that has only been observed in a very small or unrepresentative sample of subjects would be said to have little external validity in the sense that its generalizability is severely restricted. To exhibit external validity, an experiment must first demonstrate internal validity. One would hardly want to generalize an observed relationship if the validity of that relationship were itself in question. Internal validity is therefore a prerequisite for external validity.

A third—and often overlooked—form of validity is *theoretical validity*, which refers to the logical bearing of an experiment on some hypothesis or theory. Although there are occasional exceptions in the early exploratory stages of a research program, the vast majority of experiments are conducted to test a hypothesis or a theoretical prediction. No single experiment is ever crucial in such evaluations (Lakatos, 1970), but it is still important that the experimental procedures be clearly relevant to the hypothesis in question. In fact, many scientific controversies focus on what

constitute an adequate experimental test of some hypothesis (Kuhn, 1962). Moreover, researchers may argue about which hypothesis is most relevant to the current state of the art. In their analysis of this issue, Vitroff and Featheringham (1974) expand on the popular notions of Type I and Type II errors and talk about a Type III error—namely, the probability of having conducted the wrong experiment. Ideally, then, an investigator attempts to

1. design an experiment whose outcome will have clear logical bearing on some focal hypothesis (theoretical validity);
2. execute that experiment in a manner that will maximize his or her logical warrant for concluding whether an effect occurred and to what factors the effect could be attributed;

3. use procedures that will maximize the generalizability and replicability of any observed relationship.

Campbell and Stanley (1963) and Kazdin (1973) discussed some of the factors that pose potential threats to these experimental ideals. The risk of exclusion and oversimplification, the 10 most common culprits in experimental inadequacy might be considered the following:

1. selection of a theoretically irrelevant hypothesis or issue;
2. use of a subject sample that is very small or unrepresentative of the population to which generalizations are to be drawn;
3. in the case of between-subjects designs, the absence of random assignment to the various experimental conditions;
4. poor specification of the independent variable(s);
5. inadequate standardization, assessment, or description of how the independent variable was implemented;
6. inadequate control for factors other than those of immediate experimental interest;
7. inadequate replication of the cause-effect relationship (either within or between subjects);
8. poor choice, specification, or assessment of all relevant dependent variables;
9. inadequate data presentation; and
10. conclusions or interpretations that are

not logically warranted by the experimental procedures.

Since these experimental flaws are most commonly cited post hoc—as in a journal referee's comments—it might be worthwhile to briefly expand on each in the context of a hypothetical manuscript.

Anatomy of an Experimental Manuscript

For purposes of illustration, let us assume that we are evaluating a hypothetical study on the topic of whether extrinsic rewards can actually undermine a person's interest in some activity. This issue has recently become a very controversial one (cf. Deci, 1971, 1972; Lepper, Greene, & Nisbett, 1973; Levine & Fasnacht, 1974). Some attribution theorists have argued that extrinsic rewards (such as tokens) may cause a person to devalue the rewarded activity and to "infer that his actions were basically motivated by the external contingencies of the situation, rather than by any intrinsic interest in the activity itself" (Lepper et al., 1973, p. 130). Others have defended the use of extrinsic rewards in certain situations to motivate performance.

Introduction

In the introductory section of our hypothetical manuscript, it would be easy to defend the timely relevance of this issue. On the other hand, an introduction usually includes a specification of experimental hypotheses and a preface to the methods used. It is here that theoretical validity is often threatened in that a hypothesis may be oversimplified or illegitimately inferred from a parent theory. For example, in our illustration, it is hardly the case that attribution theory predicts a decline in intrinsic interest every time extrinsic reinforcement is used. Among other things, the individual's perception of the reinforcer may moderate this alleged effect (cf. Steiner, 1970). It would therefore be misleading to state the experimental hypothesis (implicitly or explicitly) as "extrinsic rewards (always) lead to a decrease in intrinsic interest."

Method

As one moves into the Method section of the article, the number of potential threats

to validity increases substantially. For example, who are the subjects? There are practical constraints on the research populations to which most of us have access, and this is understandable. We must bear in mind, however, that these constraints will take their toll in limiting the generalizability of our findings. If we study the behavior of a small group of preschoolers in a progressive academic community, our warrant for generalizations is accordingly handicapped. In a technical sense, of course, it is impossible to experiment with a "universally representative sample." We cannot hope to include a wide enough range of subjects to allow us to say that they were representative of *all* persons in *all* ways. For practical purposes one must often choose beforehand which factors are most relevant to the generalizability of our findings. Subsequent replications and extensions of a study may help clarify its external validity. The use of laboratory analogues and "nonclinical" populations also bears on external validity, and current writers appear to be divided on the conceptualization and merits of such research (cf. Bandura, 1978; Bernstein & Paul, 1971; Kazdin & Rogers, 1978).

In addition to the importance of a representative sample, most methodologists encourage the use of "random assignment" in between-subjects research. Essentially, this is intended to remove (or at least minimize) biases that might be present if one were to use a nonrandom method of deciding which subjects will experience which experimental manipulation. In our hypothetical study on intrinsic interest, for example, one might be biased in assigning children to experimental versus control groups. If the brighter children were overrepresented in one of these groups, this might constitute a threat to both internal and external validity. By using a random assignment procedure, it is hoped that auxiliary variables will be distributed evenly across the various experimental groups. With an infinitely large sample, this would, in fact, occur, but true random distributions on *all* variables are impossible when one is working with a finite sample. Once again, we face the reality that our methodological ideals can be only crudely approximated in actual practice. This is particularly apparent when one is

working with preformed groups (e.g., classrooms), which must be dealt with *en masse*. These groups often present practical impossibilities for individual random assignment, and the conclusions drawn from such endeavors must be appropriately cautious. Alternatives to random assignment include matching (yoking) and blocking (tiering), but these introduce complexities in data interpretation (Feldt, 1958).

One of the most common (and frustrating) shortcomings of an experimental article is the failure to clearly specify the independent variable(s). It would not be informative, for example, to report that "subjects were reinforced for playing with jigsaw puzzles." A colleague who might want to replicate one's efforts would be hard pressed to determine what actually took place on the basis of such a cursory and superficial description of procedure. What constituted the reinforcer? Who administered it? How was it presented? Was the experimenter always the same person? Good operational specifications of procedure are essential not only to facilitate replication but also to clearly communicate what it was that was manipulated. It is difficult to interpret an experimental outcome if one is unclear about the independent variable.

Related to this need for clear operational descriptions of procedure is the importance of assessing the presence and variability of the independent variable(s). This is particularly apparent when the presence of the independent variable is not totally in the hands of the experimenter. In our hypothetical manuscript, for example, there are at least two possible implications of the simple procedural statement that "subjects were reinforced"—namely: (a) Subjects were presented with positive stimuli, and (b) the reinforcement value of these stimuli was reflected by increases in the target behavior. This second meaning derives from the common (if problematic) definition of reinforcement as a procedure that increases the future likelihood of specified performances. In our hypothetical study, suppose the experimenters "reinforced" children with candy. One might expect little controversy over whether the children had "really" been treated in a manner relevant to the experimental hypothesis. In point of

act, however, candy might have been a relatively unrewarding stimulus under some circumstances (e.g., right after a prior classroom snack or lunch). To insure greater relevance to the hypothesis in question, one would want to (a) investigate the children's reinforcement history and/or involve them in the selection of reinforcing stimuli; (b) operationally specify how, when, and by whom these were presented; and (c) assess their effects by evaluating associated performance changes.

This digression on the assessment of independent variables is an important one and should be illustrated in a wide range of situations. Its moral would seem to be *don't take the independent variable for granted*. As a researcher, you may have gone to great pains to train therapists and/or otherwise prepare for the administration of an independent variable. These efforts may facilitate—but they cannot insure—the standard and consistent implementation of that variable in the experiment proper. Logically, it is always safest to double check and monitor the actual carrying out of experimental procedures. As the annals of history will attest, too many conclusions have been drawn on the presumed infallibility of a researcher (Barber, 1976; Mahoney, 1976). One need not lack integrity or competence to exhibit variability in the administration of even the most simple experiment, and these variations may figure prominently on the interpretation of one's data.

Inadequate specification and assessment of independent variables are certainly a common concern of journal reviewers, but they almost pale in comparison to the frequency with which a manuscript is criticized for lacking appropriate control conditions. Just as it is impossible to achieve true random assignment, it is theoretically impossible to control for all possible influences other than the variables of interest. In our hypothetical study, for example, one might include a control group that was rewarded noncontingently as an attempt to control for such factors as mood, the subjects' perception of the experimenter, and so on. Such a group might be informative, but it would hardly be sufficient in isolating potential influences. Given the infinite array of potential variables in most experiments, one

must usually decide to control for the "most likely" influences extraneous to the experimental variable(s). This obviously shades the decision of which variables are most likely to need controlling. The researcher is aided (or perhaps abetted) here by some historical consensus on the most likely "artifacts" in experimental research. At the risk of oversimplification, they are: (a) maturation (cf. Campbell & Stanley, 1963), (b) subject expectancy (cf. Wilkins, 1973), (c) experimenter bias (cf. Barber, 1976; Orne, 1962; Rosenthal, 1966; Rosenthal & Rosnow, 1969), (d) assessment (i.e., the effects of measurement alone), (e) participation (i.e., the effects of involvement in an experiment), and (f) differential attrition (i.e., selective loss of subjects from some experimental conditions).

The history of methodological refinements in the behavioral sciences is basically a history of attempts to control for such variables as these. The 12 most common experimental designs are illustrated in Table 1, which draws heavily on the early conceptualization of Campbell and Stanley (1963). In the table, a 0 refers to an observation or assessment and an X refers to some form of intervention or treatment. Reflecting some of my own biases, the designs outlined in Table 1 are listed in a crude progression of validity (i.e., with the later designs exhibiting greater internal and external validity than the earlier ones). Unfortunately, Designs 10, 11, and 12—which promise the strongest logical warrant—are among the least popular in the field.

It is also noteworthy that two of the most crucial threats to internal validity are among the most neglected in behavioral science research, that is, subject expectancies and experimenter bias. Literature ranging from the "placebo effect" to psychotherapy outcome documents the potential influence of subject expectancies in an experiment (cf. Badia, Haber, & Runyon, 1970; Kazdin & Wilcoxon, 1976; Shapiro, 1971). If a person *participates* in an experiment, his or her behavior is likely to be different than that of a nonparticipating control subject. One might say "of course—that just shows that an effect has occurred." Unfortunately, many studies suggest that simple participation in *any* experience that is presented as "treatment" may produce thera-

Table 1
The 12 Most Common Experimental Designs

Design	Symbol	Description	Comments
		Within subject	
1. Posttest only	X0	A person or group experiences a manipulation (X) (e.g., therapy), and the dependent variable is then measured (0).	Extremely weak and uninformative design; no strong conclusions can be drawn.
2. Pretest-posttest	0X0	The dependent variable is measured (0) before and after the experimental manipulation (X).	Weak design; it may be concluded that there was (or was not) a change in the dependent variable, but one cannot determine whether this change would have occurred anyway (without the experimental manipulation).
3. Reversal	0X0X0	Two separate manipulations of the independent variable are each preceded and followed by measurement of the dependent variable.	More adequate design in that it can replicate the observed effect of an experimental manipulation; conclusions are limited to the subject or group in question, however, and this design does not rule out the possible influence of factors other than the independent variable; reversal may pose practical and ethical problems in some situations.
4. Equivalent time samples	0X0X0X0X0X	An extension of the reversal design in which an independent variable is sequentially presented and removed (or otherwise manipulated) with alternating measurements of the dependent variable.	Moderately adequate design in the sense of multiple replication and possible control of some time-related factors; limitations include the possibility that the effects of a manipulation may change due to its repeated presentation and withdrawal.
5. Multiple baseline	0X000 00X00 000X0	The timing of an experimental manipulation is systematically varied across different behaviors or situations.	Moderately adequate design in that it includes replication and partial control of time-related factors; limitations vary with the specific procedures.
6. Time series	000X000	The stability of a dependent variable is measured; deviations from that stability after an experimental manipulation are used to infer causal relationship.	Somewhat controversial in terms of adequacy; limitations include failure to rule out factors that changed simultaneously with the experimental manipulation and failure to replicate.
7. Changing criterion	0X ₁ 0X ₂ 0X ₃	Similar in some respects to equivalent time samples design, except that the value of the independent variable is systematically altered; if changes in the dependent variable reliably covary with these manipulations, a causal relationship is inferred.	Moderately adequate design that shares many of the strengths and weaknesses of the equivalent time samples design; in addition, this may be problematic with some patterns of criterion change.

Table 1 (continued)

Design	Symbol	Description	Comments
		Between subject	
8. Multiple baseline	0X000 00X00 000X0	The timing of an experimental manipulation is systematically varied across persons or groups.	See multiple baseline design within subjects (5).
9. Control group	0X0 0 0	Subjects are randomly assigned to an experimental or control condition (usually groups); the independent variable is manipulated only in the experimental condition, but the dependent variable is measured (pre and post) in both.	Generally adequate design whose limitations include failure to evaluate the effects of being observed or participation in <i>any</i> experiment.
10. Solomon four group	0X0 0 0 X0 0	Subjects are randomly assigned to four conditions (usually groups), two of which will receive the experimental manipulation—one of these is tested (i.e., the dependent variable is measured) before and after the manipulation; the other is tested only after; in the two control conditions, one is tested pre and post, the other only post.	Very adequate design in that it controls for the effects of testing; limitations include failure to evaluate the contribution of participating in <i>any</i> experiment.
11. Attention and control group	0X0 0 0 X0 0 0Y0 Y0	Subjects are randomly assigned to six conditions (usually groups), four of which are identical to those in the Solomon design; the two remaining receive some form of contact or attention designed to control for simple participation in an experiment.	Very adequate design; major limitation is failure to control for subject expectancies.
12. Placebo and control group	0X0 0 0 X0 0 0Z0 Z0	Identical to Design 11, except that the two added conditions receive an experimental manipulation that has an equal degree of credibility or probable effects as the true experimental manipulation; the placebo's probable effects are evaluated by subjects.	Very adequate design; the placebo condition simultaneously controls for participation and expectancy.

peutic changes in some individuals (Shapiro, 1971). This is particularly the case when (a) the subject believes that he or she is receiving

a treatment (or experimental) variable and (b) the subject harbors positive expectancies about the probable effects of the variable in

question. Humans tend to be active generators of hypotheses, and they may often try to guess an experimenter's hypotheses as well as their group assignment. They are then more likely to be influenced by their own expectancies and what they perceive as the demands of the situation (cf. Orne, 1962).

Particularly important in research on the outcome of a psychological treatment are controls for participation and subject expectancy. This is most efficiently performed by the placebo and control group design (Design 12 in Table 1). It should be noted, however, that a "placebo" (i.e., inert) treatment only controls for subject expectancies to the extent that it is perceived (by the subject) as being equally credible and powerful as the experimental treatment(s). One cannot presume to determine these factors without involving the subjects themselves. That is, one cannot dictate the credibility of a placebo or presume that subjects will expect to benefit as much from it as from one's experimental treatment. These are empirical issues that may require pilot studies and that certainly merit assessment in each application of the placebo.

None of the designs outlined in Table 1 specify controls for experimenter bias—the artifacts that may be introduced in an experiment (consciously or otherwise) by the investigator. Although experimenter effects are seldom granted much concern in clinical research, the available evidence suggests that this complacency may be quite costly (cf. Barber, 1976; Mahoney, 1976; Rosenthal, 1966; Rosenthal & Rosnow, 1969). It is not an insult to the integrity or competency of a researcher to acknowledge that he or she may exhibit subtly human patterns of fallibility that may threaten the internal validity of an experiment. In some instances, the experimenter may bias things against his or her hypothesis, but more often than not the prejudice seems to favor rather than challenge the hypothesis in question. It is partly for this reason that *independent* replication is so important. On the other hand, we now know that scientists can independently replicate a spurious phenomenon dozens of times if they share the same expectancies (cf. Mahoney, 1976). For this reason, it is important to endeavor toward maximizing the objectivity of

the investigator(s) by such means as using experimenters who are "blind" (uninformed) about the hypotheses and/or subject conditions. Multiple raters or persons with contrary biases may also help to assess (and hopefully minimize) experimenter bias. No single method is ever completely effective in eliminating this factor, and an inability to control for experimenter bias does not mean that a study should not be conducted. It does, however, restrict one's logical warrant for confident data interpretations.

Results

The results section of a manuscript may play an important role in its fate—particularly if the results run contrary to popular assumptions or the biases of the reviewers (Mahoney, 1977). The psychology of science is not, however, our current concern. More relevant are the factors that bear on manuscript evaluation and become most salient in the Results section. These include (a) choice of dependent variables; (b) specification and assessment of dependent variables; (c) data presentation; (d) descriptive statistics; (e) inferential statistics; and (f) in treatment studies, maintenance and/or generalization of therapeutic effects.

In most cases, the choice of dependent variables may seem straightforward and uncontroversial. When concern is expressed about the outcome measures in an experiment, it most often bears on (a) the relevance of the variable to the hypothesis, (b) the failure to assess other relevant variable(s), and (c) the assessment methods used. In our hypothetical manuscript on intrinsic interest, for example, subjects' actual performance rates might appear to be a straightforward outcome measure. If the task involved jigsaw puzzles, however, one must decide whether to assess time spent in the activity, number of units completed, the "vigor" of responding, and so on. Likewise, a more comprehensive outcome assessment might be facilitated by asking subjects to rate their interest in the experimental activity.

Assessment methods constitute an ambitious topic in and of themselves, and they cannot be adequately examined here. At the risk of seeming perfunctory, I shall offer a handful of

suggestions in this regard. First, one is well advised to use *multimodal assessment*, that is, a variety of dependent variables and assessment methods that may lend consensual confidence to one's observations. Direct observations of behavior, physiological measures, subjective self-reports, unobtrusive measures, and standardized psychometric instruments may each contribute to one's confidence regarding the nature and magnitude of an experimental effect. Wherever possible, the assessment instruments should be chosen from among those that are professionally recognized as having demonstrated acceptable reliability (consistency) and relevance to the outcome in question. When direct behavioral observations are used, response categories should be clearly defined and independent interrater reliabilities should be reported. More extensive discussion of these points and some of the more popular instruments are available elsewhere (cf. Cone & Hawkins, 1977; Cronbach, 1970; Hersen & Bellack, 1976; Meehl, 1973; Mischel, 1968; Webb, Campbell, Schwartz & Sechrest, 1966).

Data presentation is another topic that could consume a volume in and of itself. Although it is usually impossible to publish one's raw data in their entirety, it is important that they be summarized in an objective and communicative fashion. It is also a scientist's obligation to make his or her raw data available to colleagues (cf. Wolin, 1962). The goal in data presentation is to be concise without sacrificing content. This can usually be accomplished by a judicious construction of figures and tables. The tables or text should include relevant descriptive statistics that report more than one measure of central tendency and indicate the range and variance of the data in question.

Use of "appropriate" inferential statistics is another controversial point, and many manuscripts encounter problems in this area. Some of the most common concerns here include

1. appropriate choice of a statistical test (e.g., parametric or nonparametric, variance or covariance analysis, etc.);
2. application of the test to the appropriate dependent variables (e.g., posttest scores versus change scores);
3. protection against probability pyramiding (Neher, 1967) through use of conservative

significance levels or appropriate post hoc comparisons;

4. failure to attend to issues of power in selecting sample size and significance level;

5. failure to consider the possible role of statistical regression in the obtained results; and

6. failure to distinguish between statistical significance and clinical significance.

The amount of emphasis placed on statistical inference is substantial, and there is likely to be vigorous debate regarding Weimer's (1977) recent demonstration that all contemporary inferential statistics are illogically founded. The logic of his proof is straightforward and may raise the ire of many a science apprentice who has been forced to develop traditional statistical proficiencies (Mahoney, 1976). I shall avoid the temptation to expand on that proof and instead suggest that we might do well to imitate the physical sciences in their rejection of null hypothesis testing as an automatic and rational arbiter of experimental results (cf. Greenwald, 1975; Meehl, 1967). Reports of percentage of outcome variance attributable to the independent variable(s) might offer a more rational and constructive alternative to our present inferential statistics. One should bear in mind, however, that this is still a minority view and that the vast majority of psychologists—and journal referees—hold few things more sacred than significance levels.

In clinical outcome evaluations, the generalization and maintenance of a therapeutic effect often constitute important factors in the professional response to a manuscript. Maintenance is a particularly salient concern, perhaps because it has been such an elusive feature in our efforts to facilitate human adjustment. Short-term therapeutic effects are not difficult to produce in some behavior disorders. Short-term improvements that can be confidently attributed to the treatment in question are much less common; enduring and attributable therapeutic improvements are a rarity indeed. It is, I think, a promising sign that contemporary journals are now demanding more refined methodologies and demonstrations of long-term maintenance.

Discussion

How one's data are interpreted depends, in part, on how they have been analyzed and

whether their implications are obvious. In the case of the latter, one can occasionally find consolation in their having passed the test of "interocular trauma," that is, when the effect or outcome is so salient that it "hits one between the eyes" (Weimer, 1977). In the absence of such a dramatic event, one must draw conclusions about (a) whether an effect was observed (e.g., therapeutic improvement), (b) whether an effect (if present) can be attributed to the independent variables, (c) how the present findings bear on the hypothesis or theory in question, and (d) how widely the present findings can be generalized. These conclusions are obviously related to the issues of internal, external, and theoretical validity.

Most researchers are well versed in the designs that will allow them to confidently assert that an experimental effect was observed (Table 1). There is much more room for interpretive license (and error) when it comes to asserting that the independent variables were responsible for such an effect. It is here that adequate controls and protections of internal validity are most relevant. In our hypothetical study on intrinsic interest, for example, one could conclude very little from a control group design (see Design 9, Table 1) due to the plethora of competing variables (e.g., access to the target task, boredom, etc.) Errors or excessive license in causal attribution are a common concern of journal reviewers, as are overgeneralizations (cf. Maher, 1978). Interestingly, another form of logical error that is seldom noted may be one of the most common, and its effect may be among the more pernicious.

When the experimental data are consistent with one's hypothesis (and the null hypothesis has been rejected if one has been stated), it is very common for the researcher to conclude his or her article with a *confirmation claim*. In its essence, a confirmation claim is an assertion that the truth value of a proposition or hypothesis has been strengthened, proven, or otherwise augmented. This is unfortunate and misleading, because *no empirical hypothesis can ever be confirmed* (Mahoney, 1976; Weimer, 1977). Confirmation refers to a form of logical inference in which a true premise warrants a true conclusion. Consider, for example, the hypothesis that "if theory A

is true, then observation X will occur." Logically, the confirmation of this hypothesis requires that one be told that the premise is true, after which one could rationally infer that the conclusion is also true (i.e., that observation X will, indeed, occur). This is obviously of little use to the scientist, since the truth value of the premise is most often the point at issue. Philosopher Karl Popper has noted this for years, and yet scientists persist in their claims of confirmation (cf. Popper, 1959, 1963, 1972).

When a researcher reports findings that are consistent with his or her hypothesis and then argues that these data "support" the hypothesis, a serious logical error has been committed. Technically, the error is called *affirming the consequent*. It consists of making the illogical inference that a true conclusion implies a true premise. In my earlier illustration (in which theory A predicted observation X), the occurrence of observation X has absolutely no logical bearing on the truth value of the premise (theory A). The observation of not X would, however, bear clear logical implications for the theory. As we all learned in elementary algebra, it is only false conclusions that bear conclusively on a premise (i.e., they can falsify it). This was, in fact, the cornerstone of Popper's philosophy of falsificationism. Counterintuitive as it may seem, negative results and predictive failures have far-reaching logical implications, and positive results (successful predictions) have comparatively little information content. This is, of course, contrary to the popular practice of selectively publishing positive results manuscripts and emphasizing these successes more heavily in literature reviews (Meehl, 1967). The epistemological costs of this practice are themselves quite distressing (cf. Mahoney, 1976; Smart, 1964).

So what does one do with "positive results," that is, an experimental outcome that is consistent with the hypothesis in question? How are these to be interpreted? In a word, cautiously. One can legitimately note their consistency with the hypothesis or state that they *corroborate* the hypothesis. Loosely speaking, corroboration refers to the act of having survived falsification (Popper, 1972). To say that a hypothesis has been "corroborated" is

simply to state that it has been tested and has (tentatively) survived. This is very different, of course, from implying that it has been strengthened. No matter how many times a hypothesis survives falsification, it can never be legitimately said to have been "confirmed" or "supported." As corroborative instances accumulate, most of us experience a *psychological* increase in our confidence regarding the hypothesis in question. This subjective phenomenon should not, however, be confused with the logical warrant of the situation. As much as we scientists may like to view ourselves as rational creatures, there is a clear and significant difference between our rare use of logic and more common use of psycho-logic.

Concluding Remarks

It has not been my intent to overemphasize the shortcomings of science or to criticize our attempts to use experimental methods in the refinement of therapeutic interaction. We must recognize, however, that all of our scientific efforts fall along a continuum of fallibility. There is no investigation that can be *totally* lacking in its potential informativeness, nor will there ever be one that is perfect in its attainment of internal, external, and theoretical validity. Our goals, then, should be to strive toward conducting the least fallible inquiries, to cautiously interpret our experiments in accordance with their logical warrant, and to guard against the paralysis of complacency regarding the adequacy of current research methods. One of the most pervasive aspects of scientific endeavors has been their exploratory character, and we are, I think, well advised to explore refinements in both our methods as well as our technical knowledge. As illustrated in the cases of inferential statistics and confirmation claims, these explorations may require some psychologically unsettling changes—from old "accepted" procedures to newer alternatives. As such, they may require a reappraisal of long-revered assumptions about the best way to do scientific inquiry. In the final analysis, however, the prospect of progress will hopefully overshadow the inconvenience it may demand. We can, after all, hardly expect to "grow" if we are unwilling to change.

References

- Badia, P., Haber, A., & Runyon, R. P. (Eds.). *Research problems in psychology*. New York: Addison-Wesley, 1970.
- Bandura, A. On paradigms and recycled ideologies. *Cognitive Therapy and Research*, 1978, 2, 79-103.
- Barber, T. X. *Pitfalls in human research: Ten pivotal points*. Elmsford, N.Y.: Pergamon Press, 1976.
- Bernstein, D. A., & Paul, G. L. Some comments on therapy analogue research with small animal "phobias." *Journal of Behavior Therapy and Experimental Psychiatry*, 1971, 2, 225-237.
- Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1963.
- Cone, J. D., & Hawkins, R. *Behavioral assessment in clinical psychology*. New York: Brunner/Mazel, 1977.
- Cronbach, L. J. *Essentials of psychological testing*. New York: Harper & Row, 1970.
- Deci, E. L. Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, 1971, 18, 105-115.
- Deci, E. L. Intrinsic motivation, extrinsic reinforcement, and inequity. *Journal of Personality and Social Psychology*, 1972, 22, 113-130.
- Feldt, L. S. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, 1958, 23, 335-353.
- Greenwald, A. G. Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 1975, 82, 1-20.
- Greenwald, A. G. Within-subject designs: To use or not to use? *Psychological Bulletin*, 1976, 83, 314-320.
- Hersen, M., & Barlow, D. H. *Single case experimental designs*. New York: Pergamon Press, 1976.
- Hersen, M., & Bellack, A. S. *Behavioral assessment: A practical handbook*. New York: Pergamon Press, 1976.
- Hook, S. (Ed.). *Determinism and freedom in the age of modern science*. New York: Collier, 1958.
- Kazdin, A. E. Methodological and assessment considerations in evaluating reinforcement programs in applied settings. *Journal of Applied Behavior Analysis*, 1973, 6, 517-531.
- Kazdin, A. E., & Rogers, T. On paradigms and recycled ideologies: Analogue research revisited. *Cognitive Therapy and Research*, 1978, 2, 105-117.
- Kazdin, A. E., & Wilcoxon, L. A. Systematic desensitization and nonspecific treatment effects: A methodological evaluation. *Psychological Bulletin*, 1976, 83, 729-758.
- Kuhn, T. S. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1962.
- Lakatos, I. Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, England: Cambridge University Press, 1970.

- Lakatos, I., & Musgrave, A. (Eds.). *Criticism and the growth of knowledge*. Cambridge, England: Cambridge University Press, 1970.
- Lepper, M. R., Greene, D., & Nisbett, R. E. Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, 1973, 28, 129-137.
- Levine, F. M., & Fasnacht, G. Token rewards may lead to token learning. *American Psychologist*, 1974, 29, 814-820.
- Maher, B. A. Stimulus sampling in clinical research: Representative design reviewed. *Journal of Consulting and Clinical Psychology*, 1978, 46, 643-647.
- Mahoney, M. J. *Scientist as subject*. Cambridge, Mass.: Ballinger, 1976.
- Mahoney, M. J. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, 1977, 1, 161-175.
- Meehl, P. E. Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 1967, 34, 103-115.
- Meehl, P. E. *Psychodiagnosis*. Minneapolis: University of Minnesota Press, 1973.
- Mischel, W. *Personality and assessment*. New York: Wiley, 1968.
- Mitroff, I. I. *The subjective side of science*. New York: Elsevier, 1974.
- Mitroff, I. I., & Featheringham, T. R. On systematic problem solving and the error of the third kind. *Behavioral Science*, 1974, 19, 383-393.
- Neher, A. Probability pyramiding, research error and the need for independent replication. *Psychological Record*, 1967, 17, 257-262.
- Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 1962, 17, 776-783.
- Paul, G. L. Behavior modification research: Design and tactics. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- Popper, K. R. *The logic of scientific discovery*. New York: Harper, 1959.
- Popper, K. R. *Conjectures and refutations*. New York: Harper, 1963.
- Popper, K. R. *Objective knowledge: An evolutionary approach*. London: Oxford University Press, 1972.
- Rosenthal, R. *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts, 1966.
- Rosenthal, R., & Rosnow, R. L. (Eds.). *Artifact in behavioral research*. New York: Academic Press, 1969.
- Shapiro, A. K. Placebo effects in medicine, psychotherapy, and psychoanalysis. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change*. New York: Wiley, 1971.
- Sidman, M. *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books, 1960.
- Smart, R. The importance of negative results in psychological research. *Canadian Psychologist*, 1964, 5, 225-232.
- Steiner, I. D. Perceived freedom. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. New York: Academic Press, 1970.
- Thoresen, C. E. *Let's get intensive: Studying change over time*. Englewood Cliffs, N.J.: Prentice-Hall, in press.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally, 1966.
- Weimer, W. B. *Notes on methodology*. Hillsdale, N.J.: Erlbaum, 1977.
- Wilkins, W. Expectancy of therapeutic gain: An empirical and conceptual critique. *Journal of Consulting and Clinical Psychology*, 1973, 40, 69-77.
- Wolin, L. Responsibility for raw data. *American Psychologist*, 1962, 17, 657-658.

Received September 21, 1977

Evaluating the Generality of Findings in Analogue Therapy Research

Alan E. Kazdin
Pennsylvania State University

Analogue research in psychotherapy and behavior therapy has proliferated in recent years. The value of analogue research is that it allows analytic and well-controlled research to address questions that often are prohibitive or impractical to evaluate in clinical situations. The major source of controversy about the value of analogue studies is their external validity, that is, the extent to which the results can be generalized to the clinical situation. Much of the controversy stems from conceptualizing therapy investigations discretely as either analogue or clinical research. However, all treatment research is an analogue of the situation to which an investigator wishes to generalize. Thus, the main question is the extent to which an investigation is an analogue of the clinical situation. An investigation can vary from the clinical situation along several dimensions, such as the target problem, the clients studied, the manner of client recruitment, and others. It is often assumed that the greater the similarity of an intervention to the clinical situation, the more likely the findings will be generalizable to the clinical situation. The present article questions this assumption and suggests an empirical method for evaluating the generality of therapy research to the clinical situation.

Investigation of treatment outcome has been and continues to be a major research topic in psychotherapy and behavior therapy (e.g., Bergin, 1971; Eysenck, 1952; Kazdin & Wilson, 1978; Luborsky, Singer, & Luborsky, 1975; Meltzoff & Kornreich, 1970; Paul, 1969; Rachman, 1971; Sloane, Staples, Cristol, Yorkston, & Whipple, 1975). The goals of therapy outcome research usually are to determine the efficacy of a given treatment, to evaluate the relative effectiveness of different

treatments, and to assess the components of treatment that are responsible for change for a particular treatment problem and client population. The most direct means of achieving these goals is to study the therapy techniques and populations of interest directly in a clinical setting. The requirements for conducting clinical research appear relatively straightforward. Clients who seek treatment for a given problem at a treatment facility can be assigned to different treatment or control groups, depending on the research questions and desiderata of the experimental design. Treatment can be administered by trained therapists and evaluated by assessment of therapeutic change on multiple measures of the clients' problems.

This article was written while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, California. Fellowship support was obtained from the Foundations Fund for Research in Psychiatry, National Institute of Mental Health Grant 1-T32-MH14581 and the Spencer Foundation, which the author gratefully acknowledges.

The author is grateful for the extensive discussions and specific comments on the article provided by W. Stewart Agras, Nathan H. Azrin, Alexander George, Walter Mischel, Stanley Rachman, and G. Terence Wilson, who also were Fellows at the Center for Advanced Study in the Behavioral Sciences.

Send requests for reprints to Alan E. Kazdin, Department of Psychology, Pennsylvania State University, University Park, Pennsylvania 16802.

Although the basic requirements for outcome research might be highlighted in a straightforward fashion, meeting these requirements in practice presents nearly insurmountable obstacles for the researcher. Because of the diverse practical and ethical obstacles of clinical research and the complexity of treatment, a great deal of research has been

conducted in situations analogous to those available in the clinic. Research that evaluates treatment under conditions that only resemble or approximate the clinical situation has been referred to as "analogue research." An analogue study usually focuses on a carefully defined research question under well controlled conditions. The purpose of this investigation is to illuminate a particular process or to study an intervention that may be of importance in actual treatment.

An important issue in the psychotherapy and behavior therapy literature is the extent to which analogue studies contribute to understanding therapeutic processes and outcome in clinical settings. It is generally acknowledged that analogue studies are essential and that they provide an important link in extrapolating from basic laboratory research to therapy. However, even in the most well-designed analogue investigations, there is a lingering concern about the external validity of the results, that is, the generalizability of the findings to the clinical situation.

Whether analogue studies provide a basis for drawing conclusions about the clinical efficacy of a given treatment technique is a matter of dispute. The purpose of the present article is neither to defend nor to criticize the use of analogue research. Indeed, a basic tenet that will be advanced is that experimental treatment research invariably is analogue research. The purpose of the article is threefold. First, treatment research is conceptualized in such a way as to question the utility of the notion of analogue research as a distinct category of therapy studies. Second, dimensions for determining the extent to which a given study is an analogue of clinical treatment are examined. Finally, and most importantly, a way of studying directly the relationship between analogue and treatment studies is suggested. Hopefully viewing the analogue nature of treatment research as a matter of degree and suggesting a method for investigating generality of the findings of treatment research will stimulate an empirical research in questions of external validity.

Utility of Analogue Studies

Conducting therapy research in the clinical setting in which treatment normally is pro-

vided presents several obstacles. The obstacles encompass practical difficulties of undertaking and completing clinical research, as well as ethical considerations that militate against addressing select research questions. The practical problems include diverse aspects of the investigation. Initially, in most treatment studies, it is difficult to obtain a sufficient number of clients with the same or similar problems to meet the requirements for an experiment. Similarity among the clients' problems is required so that measures of their personality or behavior can be compared along the same response dimension (e.g., anxiety, depression, or another dimension) to evaluate the effects of treatment.

Even if clients with similar problems can be included in a given study, often there are constraints in assigning clients to specific forms of treatment. For example, in inpatient treatment, administrative demands sometimes dictate that the more severe patients be given the treatment that is more likely to alter behavior than is some other treatment. In outpatient treatment, clients are not easily assigned to different variations of therapy, because they may seek and express preference for specific treatments. Thus, randomization of assignment, an essential ingredient in most experimentation, may not be feasible.

Another problem in working with clinical populations is that it may be difficult to control or remove the influence of competing factors that can affect therapy results. For example, in outpatient psychotherapy, clients may receive additional experiences (e.g., encounter group or marathon sessions) or counsel (e.g., from a physician) while participating in a given treatment. In inpatient treatment with psychiatric patients, it may be difficult to control ongoing treatments (e.g., drugs) and other factors (e.g., release from the hospital) that can influence treatment or its evaluation.

Aside from obstacles in selecting clients, it is often difficult to obtain therapists who will engage in treatment and meet the demands of research. Initially, there is little incentive for most practicing clinicians to participate in an experiment. Also, it is difficult to recruit clinicians who are homogenous in training and experience. Finally, most experienced therapists are interested in conducting treatment in their own way rather than following the some-

what regimented or standardized procedures that may be needed to evaluate a specific treatment.

Ethical considerations also make clinical investigation of certain questions very difficult (Stuart, 1973). Many important questions about therapy require control groups that withhold specific aspects of treatment that may be crucial for behavior change. To subject clients to "control" conditions that have a low probability of effecting behavior change could violate the professional commitment to treatment. Also, the possible ineffectiveness of a control procedure may lead to relatively high levels of attrition. The problem of assigning clients to control conditions also is evident in cases in which the effects of therapy are evaluated over a protracted period against a group that never receives treatment. A no-treatment or waiting-list control group that may never receive treatment or that receives treatment after a long delay (after they have served their purpose as no-treatment subjects) often is difficult to implement for clients whose problems are in a crisis state when they seek treatment.

Because of the practical and ethical problems highlighted here, therapy research on patient populations with trained therapists in clinical settings is relatively rare. Much of what is known about therapeutic processes and behavior change is learned from analogue studies. The analogue studies are investigations of circumscribed therapeutic processes or problems in a well-controlled laboratory situation. There are many advantages of laboratory investigations of clinical phenomena (Bernstein & Paul, 1971; Levis, 1970; Paul, 1970). As a general statement, laboratory investigations of therapy allow the investigator to control the conditions of experimentation to a much greater extent than do clinical investigations. This careful control allows the investigator to minimize sources of variance that might obscure an effect of treatment in situations in which several parameters are free to vary. Thus, the subjects who receive treatment can be selected because of their similarity in the type of target problem, the severity of the problem, and subject and demographic variables that might contribute to the variability of treatment effects. Aspects of treatment can be carefully controlled in analogue

research that otherwise would vary considerably in treatment. For example, therapists can be selected because of their homogeneity with respect to experience, age, training, and other factors. Also, the number of sessions, duration of treatment, and specific tasks in each treatment session can be held constant across subjects and groups. All of these characteristics of laboratory investigations will minimize variability among subjects and therapists and increase the power of the test of a given intervention.

Laboratory investigations also allow the use of different control groups that might not otherwise be available. In laboratory research, specific ingredients of treatment, or indeed treatment itself, can be more readily withheld than in the clinical situation. Freedom in providing or withholding specific ingredients of treatment greatly expands the range of questions that can be asked about therapy. This perhaps is the greatest advantage of laboratory investigations of therapy.

The procedures used in laboratory investigations probably are more easily replicated than are those used in clinical investigations. In laboratory research, investigators can more carefully control and specify the parameters of treatment administration than in clinical situations. Subsequent investigators who wish to replicate the original laboratory study may have relatively explicit guidelines to follow because many parameters of treatment were well specified originally. In the clinical situation, all sorts of conditions must be allowed to vary and may not be easily repeated in subsequent research.

The superiority in control that laboratory research affords relative to clinical research is a function of the priorities of these different research methods. In laboratory research, which usually contains volunteer subjects rather than patients seeking treatment, the higher priority is given to the demands of experimentation. Sacrifices of treatment can be made to complete the requirements of the design. For example, the treatment can be standardized across clients rather than take into account individual differences with respect to the problem. In contrast, in clinical research with clients who have sought treatment, the higher priority is their improvement. Rigors of the laboratory may have to be sacrificed

to meet the priority of treatment. For example, the therapeutic material covered during the therapy session and duration of treatment may be allowed to vary widely so that clients receive sufficient treatment to affect the desired change.

To meet the rigors of experimentation, analogue research can be conducted in circumstances that resemble the clinical situations in varying degrees. The use of analogue studies has been widely discussed (e.g., Bernstein & Paul, 1971; Cooper, Furst, & Bridger, 1969; Cowen, 1961; Goldstein & Dean, 1966; Heller, 1963, 1971; Levis, 1970). These discussions have addressed the extent to which analogue studies are useful in understanding treatment in the actual clinical situation, the limitations of analogue research, and areas in which particular care may need to be taken in substituting experimental rigor for clinical relevance.

Conceptualization of Analogue Research

As usually discussed in the clinical literature, therapy investigations are looked upon as analogue or nonanalogue research. Yet, this may not be the most fruitful way to view clinically relevant research. Dichotomizing clinical research in this fashion may obscure interpretation of research findings in different ways. First, the distinction tends to overlook the inherent limitations and the "analogue" nature of all clinical research including those studies conducted in clinical settings with patient populations. Second, and more important for present purposes, the categorization does not provide clear guidelines to distinguish among analogue studies. Presumably, generality of the results from all analogue studies to the clinical situation is not equal and partially depends on specific characteristics of the individual study.

Virtually all psychological experimentation with human subjects is analogue research insofar as it constructs a situation in which a particular phenomenon can be studied. The phenomenon is selected as an approximation of the phenomenon in the nonexperimental situation. The experimental version of the phenomenon may resemble the naturally occurring phenomenon in varying degrees, but

in an important sense it is only an analogue.¹ It is assumed, and often supported, that the versions of the phenomenon in the experimental situation and "real world" are in important ways the same and come under identical principles and laws.

In human experimentation, the differences between experimental and nonexperimental situations may lead to important differences in subject behavior. These differences can readily alter the phenomenon that is being studied and the generality from the experimental to nonexperimental setting. For example, experiments with humans place subjects in a special situation and provide some form of intervention. Subjects may complete specific response measures before and after the intervention to assess change in particular responses. However, participating in an experiment or a contrived arrangement that is not normally encountered in everyday life makes the experiment an analogue of the situation to which one may wish to generalize the results. Indeed, such phenomena as demand characteristics of the experimental situation (Orne, 1969), subject roles (Weber & Cook, 1972), and pretest sensitization (Lana, 1969) are examples of the influences of experimental arrangements on behavior.

Even when the experimental situation is not contrived and does not differ from the situation in which the client normally behaves, other features of human experimentation contribute to its analogue nature. In an experiment, psychological measures (e.g., self-report inventories, behavioral checklists, physiological responsiveness) are used to assess the relationship between an intervention (e.g., treatment) and behavior change. Yet, it is not the response on the psychological measurement device per se that is of interest, but rather it is the construct that is assumed to be represented by the measure. For example, in clinical research in an actual treatment setting, a client's anxiety might be assessed with various devices. It is not the client's change on these

¹ Not all human psychological research is an analogue of the situation to which the investigator wishes to generalize. For example, research on the influence of the experimental situation itself on behavior is a direct test of the subject matter of interest.

measures per se that is of direct interest, at least to the client. These measures are of interest, because they may reflect or relate to changes occurring in the natural environment under ordinary circumstances. Insofar as experimental research uses measures that are considered to reflect or resemble the responses of clinical interest in the natural setting, it is an analogue of the situation of direct interest.

Viewing almost all psychological research with humans as analogues of situations to which one would like to generalize has important implications for conceptualizing treatment research. Initially, it is essential to keep in mind that investigators are interested in extrapolating the findings to some area, problem, or setting that is not studied directly. There is always the possibility that the extrapolation is not accurate in accounting for the natural phenomenon. The difference between the natural and experimental situation precisely along these (usually unknown) dimensions might lead to different findings for a given variable. A second implication is that it may not be useful to speak of analogue versus nonanalogue research. Rather, research can be viewed on the basis of the extent to which it resembles the situation to which one wishes to generalize the experimental findings. There are several dimensions along which therapy research may vary from the clinical situation that will be discussed below.

Experimental research usually is removed in some way from the situation to which one wishes to generalize. This by itself does not necessarily limit the extent to which results can be generalized. The results of studies often can be widely generalized across dimensions that may seem more discrepant than the leap from laboratory and clinical investigations in the area of therapy. For example, most findings in the psychology of learning have been obtained in experiments with infrahuman species. Specific learning paradigms and principles have had generality to human behavior despite differences between humans and infrahumans and the importance of considering uniquely human characteristics. Similarly, outside of psychology, infrahuman species have been used to study genetics. The laws that have been uncovered have applied to hereditary transmission in humans, even

though the population often studied (e.g., fruit fly) bears little resemblance to humans. As a general point, research conducted in the context of the laboratory under conditions that seem very distant from the one to which a researcher wishes to generalize may produce laws or generate theories of great relevance to areas that are not directly studied. Thus, the fact that a study does not focus on a problem and population of direct interest is not always critical.

Dimensions for Evaluating Therapy Analogues

In psychotherapy and behavior therapy research, the question of interest is not whether a particular study is an analogue. As noted earlier, experimentation even in the clinical rather than laboratory situation is removed from the actual phenomenon and situation to which one might wish to generalize. Thus, in any given therapy study, the question is to what extent the conditions resemble the situation to which the investigator wishes to generalize. This is a complex question, because there are a large number of dimensions along which studies may vary. Each study may be evaluated in terms of its standing on these dimensions to determine the degree to which the study approaches the clinical situation. Major dimensions that can be used to evaluate a study and the extent to which it is an analogue of the clinical situation include the target problem; the population; the manner of client recruitment; the therapists; the selection, set, and setting of treatment; the variation of treatment used; and the assessment procedures used.

Target problem. A major concern about the value of analogue studies, particularly in recent years, pertains to the target problem (Cooper et al., 1969; Levis, 1970). The issue is whether subjects in a laboratory study evince the same kinds of problems and the same magnitude of severity for a given problem as do the clients for whom treatment is usually provided. For example, in behavior therapy, a vast literature has appeared on the treatment of mild or subphobic levels of fear. Subjects typically are college students who express some level of fear on a questionnaire and

who refuse to approach a feared stimulus (e.g., harmless snake). The level of fear may be much less potent than one would encounter with a clinical population. Indeed, the percentage of subjects who meet the definition of "fearful" in laboratory studies vastly exceeds the percentage of debilitating fear for the same problem in the general population (cf. Bernstein & Paul, 1971). In any case, selecting mildly fearful college students raises the question of the generality of treatment effects to clinical patients whose fears (or other problems) are more intense. Certainly, the concern is well taken, because it should be easier to alleviate mild problems than it is to alleviate severe problems.

Aside from severity of a given problem, the specific behavior studied in therapy research may vary in topography and qualitative characteristics from clinical problems to which the results might be generalized. Some problems that have been frequently studied in laboratory research may bear little resemblance to clinical problems. For example, most behavior therapy techniques designed to alleviate anxiety have been evaluated with clients who fear small animals. Recent research has revealed that fears of small animals in college students typically habituate more quickly in response to the anxiety-provoking stimulus and are more influenced by suggestion than are fears in social situations (e.g., heterosexual anxiety, speech anxiety) (Borkovec, Stone, O'Brien, & Kaloupek, 1974; Borkovec, Wall, & Stone, 1974; Singerman, Borkovec, & Baron, 1976). Results from studies focusing on behaviors that do not rapidly habituate and are less amenable to demand influences are more likely to be generalizable to clinical problems that presumably share these characteristics (Borkovec & O'Brien, 1976).

The extent to which the target problem in a therapy investigation resembles clinical problems is a matter of degree. To obtain target problems that increasingly resemble the clinical problem, a researcher can screen for severity of the problem in a nonclinical population so that the subjects included in treatment experience some deleterious effects of their behavior in everyday life. Stringent criteria for subject selection make extrapolation of the findings to a clinical population much

more plausible. For example, Lang (1968), who conducted the earliest analogue studies of desensitization, indicated that only 1%-2% of college students will evince intense fear if rigidly screened on behavioral, interview, and self-report criteria. When treatment alters intense fears of college students who are stringently selected, it is not unreasonable to extend the findings to a clinical population.

Population. The characteristics of the population (subjects or clients) that receives treatment contribute to the degree to which an investigation is an analogue of the clinical situation. The population dimension encompasses all features that might distinguish subjects in an experiment from those for whom the results might be generalized. For example, in analogue fear studies common in behavior therapy, subjects often are college students. The study is an analogue of clinical treatment in the sense that college students as a group may differ on several dimensions from clinical populations typically treated for anxiety disorders in such characteristics as level of education, age, socioeconomic status, and occupation (or lack of one). The differences may bear directly on the effects of particular therapy procedures. For example, the age of the clients and situational characteristics of their lives may bear on the kinds of problems that they bring for treatment. The influences also may covary with the receptivity of individuals to professional treatment and to their expectancies for cure. The generality of findings from an investigation of treatment to a clinical situation may partially depend on the similarity of (nonproblem) characteristics of subjects to clients who usually come for treatment.

Manner of recruitment. In most treatment settings, clients solicit a therapist on their own or are referred by a particular source (e.g., physician, relatives). The manner of obtaining clients in therapy or institutional treatment differs markedly from the manner in which subjects are obtained in an experimental investigation of treatment. In many treatment studies, college students voluntarily participate in treatment because of small rewards toward college course credit, completion of a course option, or interest in learning about a specific form of therapy. Presumably, the in-

incentives for patients who directly seek out treatment through the usual channels differ drastically from the incentives of college students who may be less concerned with being "cured" of a particular problem. The volunteer patient and the mildly coerced college student who is not very much interested in the treatment or problem for which it is applied probably represent end points on the continuum of the extent to which a given therapy study is an analogue of a clinical treatment, at least on the dimension of manner of recruiting subjects. Perhaps, more toward the middle would be subjects who are solicited from newspaper, television, or radio advertisements that make available treatment to individuals in a community situation. This form of recruitment still solicits subjects, unlike most therapy situations. However, it is likely to uncover clients who are interested in treatment even though they would not have sought treatment without the advertisement.

The manner of recruitment also relates to the factors that keep a client in therapy. In clinical treatment, clients usually are free to terminate therapy and, indeed, readily do so without penalty. In treatment research in which attrition can be devastating to the results, specific contingencies can be invoked to retain subjects. For college students who participate in experimental treatment research, either explicit or implicit contingencies operate, such as withholding credit toward course completion until a project is terminated. With clients who are solicited for treatment, refundable deposits may be collected and returned later only if treatment has been completed. The threat of losing a refundable deposit significantly thwarts attrition (Hagen, Foreyt, & Durham, 1976). Clients in treatment usually pay for therapy and hence exert ultimate control over selection and termination of therapy. Although paying for therapy per se has not been found to influence therapeutic outcome, it does covary with other variables such as diagnosis and socioeconomic status that do (Pope, Geller, & Wilkinson, 1975). In any case, the different contingencies that maintain clients in treatment may influence the generality of findings across laboratory and clinical settings.

Therapists. In the clinical situation, an

experienced therapist usually provides treatment. On the other hand, experienced full-time therapists are infrequently used in most therapy research, although there are outstanding exceptions (e.g., Orlinsky & Howard, 1975; Paul, 1966; Sloane et al., 1975). Typically, graduate students with some interest or experience in clinical applications serve as therapists. The similarity of students to therapists who apply therapy techniques in clinical situations is a dimension that may determine the extent to which the results can be generalized from research to practice.

The therapist dimension is complex, because it may include multiple differences in the characteristics of individuals who provide treatment (see Meltzoff & Kornreich, 1970; Truax & Mitchell, 1971). Initially, student therapists and professional clinicians may differ on such variables as experience, training, age, and other characteristics. In addition, they may behave differently toward their clients. Therapists may engage in certain types of conversation, hold greater expectations for improvement, and have a stronger commitment to patient cure than do students serving as therapists. The generality of treatment effects in a therapy study to the clinical situation may depend in part on who serves as therapists.

Selection, set, and setting of treatment. In clinical work, clients seek a particular treatment that they believe will be effective. Independently of how well the technique in fact helps people, the set with which people arrive at treatment is that they will receive a bona fide treatment designed to alter their behavior. In addition, the client may have heard about the particular treatment and therapist with whom he or she will have contact. The assignment of clients to treatment may be selective, since individuals exert some choice over treatment. If treatment is not viewed favorably by a client after having attended some of the sessions, choice can be exerted by leaving and attending another treatment. In experimental therapy research, clients often are assigned to treatment and are not given a choice over the treatment or therapist. The element of choosing or seeking out a particular treatment may influence the generality of results obtained in therapy

research. Indeed, therapy investigations have indicated that being able to choose treatment is highly related to therapeutic outcome, independently of the specific technique that is actually received (Devine & Fernald, 1973; Gordon, 1976).

The client's set about treatment, perhaps related to choice and selection mentioned above, may contribute to the generality of results from treatment research to clinical applications. The set about receiving an effective treatment is likely to be enhanced by the aura of the treatment facility or office and by the convincing description of therapy provided by the therapist (Frank, 1973). Interestingly, the set of the client when arriving at treatment and the manner of presenting treatment may differ markedly across research and clinical settings. In the research setting, the client may not have the initial set that an effective treatment will be provided. Also, the actual setting (e.g., university building rather than a clinic) may not foster the set for receiving help with one's problem. In addition, instructions on the part of the investigator may indicate that the procedure to be used is an "experimental" treatment.

In recent years, ethical standards for research increasingly have stressed that investigators should explicitly convey the nature of the treatment, whether it is experimental in nature, and the possible benefits and risks. The procurement of informed consent in the context of experimental clinical research may alter to some extent the manner or context in which treatment is provided. In ordinary therapy, patients usually are not told that treatment is experimental but rather are given a convincing rationale about its efficacy and illustrations of other cases. In contrast, "experimentation" usually conveys the exploratory nature and tentative effects of treatment. Generalizing the results from research to the clinical situation depends in part on the similarity of the manner in which treatment is presented and the sets of the individuals who participate in treatment.

The dependence of treatment effects on the context in which treatment is presented was demonstrated clearly by Bootzin (Note 1) in a laboratory study designed to alter fear

of rats in college students. Bootzin compared live modeling (observing someone perform a fearless response, touching and holding the rat) versus live modeling plus participation (observing plus actually engaging in touching and holding of the rat). For one half of the subjects in each of these groups, the experience was called a "treatment" demonstration; for others, the experience was labeled as a maze learning demonstration (since the rats eventually were run through a maze at the end of the session). Interestingly, live modeling plus participation decreased fear of rats (on an assessment battery administered outside of the context of this experiment) whether or not subjects were told that the demonstration was "treatment for fear." In contrast, subjects who received live modeling showed a reduction of fear only if they served in the group that specifically identified the procedure as treatment. Thus, the effectiveness of modeling depended on being identified as a form of treatment, whereas the effectiveness of modeling plus participation did not.

The general point here is that the context in which the procedures are presented may differ across studies and distinguish the specific effects of treatments. This creates some obvious problems for research that often cannot present the treatment program with the same amount of zeal that might be the case in actual clinical practice.

Variations of treatment. The extent to which the results of a therapy study can be generalized to the clinical situation depends on the degree to which the actual treatment is varied across research and clinical applications. Occasionally, changes are made in specific components of treatment to increase precision in research, to control variability across subjects, or to examine a component of treatment that might be difficult to study without the variation. Even though some changes may pertain to ancillary features of treatment, others might represent significant departures.

Several components might vary across the research and clinical versions of a given treatment. For example, research applications may hold constant the number of treatment sessions across groups; the duration of treatment sessions; the material, tasks, or topics dis-

cussed within sessions; and so on. These dimensions usually are not specified in clinical applications, and they vary freely. Yet, in research the dimensions may be rigidly specified to ensure that they do not systematically vary across treatment conditions or subjects within a condition.

Some of the components may be more important than others in standardizing treatment to make it well suited to the demands of research. For example, portions of treatment in an experiment may be prerecorded and presented by tape to the client. Perhaps, more seriously, standardization of an aspect of treatment may change a crucial ingredient of therapy that would vary treatment efficacy. For example, in systematic desensitization, clients imagine specific scenes to overcome anxiety. These scenes are hierarchically arranged and individually tailored to a client's problem. In clinical research in which specific components of treatment are standardized across groups, individuals may receive the same hierarchy of items based on the assumption that it will be germane and approximately suited to their individual needs. Yet, technically, this is a violation of the original requirement of treatment. Traversing items that are not designed for a given individual violate the goal of minimizing anxiety throughout treatment and of ensuring that relaxation is the dominating response during imagery.

Perhaps an even more extreme departure from clinical treatment is the use of slides presented to the subject as part of desensitization rather than having the client imagine the scenes (e.g., Brown, 1973; Wilson, 1973). While the client is relaxed, the slides are to be viewed as an analogue version of desensitization in which imagery is paired with relaxation. The use of slides represents a major procedural variation of treatment. Yet, this variation is consistent with the original theoretical rationale, which suggested that fear stimuli, however presented, and an incompatible response need to be paired (Wolpe, 1958). In general, many sources of departure exist that could decrease the resemblance of research and clinical versions of treatment.

Assessment procedures. The extent to

which a therapy investigation resembles the clinical situation is influenced by diverse aspects of the assessment procedures including the reactivity of assessment, the setting in which assessment is conducted, and the precise measures used. To begin with, the client's problem behavior usually is measured in situations in which the client is fully aware of the purposes of assessment. Reactivity or awareness of the assessment process itself may influence the type of information obtained about the client's behavior. Once aware of assessment, the client's responses might be influenced by response sets, biases, and demands of the situation that cue certain types of responses (Bernstein, 1973). The goal of therapy is to change behavior in the client's everyday performance that is not under the scrutiny of a psychologist. Demonstrations of behavior change on reactive measures that are under the scrutiny of a psychologist may only be a distant approximation of the relevant changes that the client wishes to achieve. The extent to which a study measures everyday performance in social situations or approximations of that performance may determine the generality of results to clinical situations.

Part of the artificiality of assessment derives from the setting in which the effects of treatment are evaluated and the specific assessment devices used. Consider first the setting and the contribution it makes to evaluating behavior change. Several studies have demonstrated the effect of the setting in which behavior is assessed by having subjects perform in the context of a "clinic" versus a laboratory setting (Bernstein, 1973; Bernstein & Nietzel, 1973, 1974). The actual physical setting is not the crucial ingredient. Rather, the clients are told that assessment is taking place in a clinical setting as part of a treatment program or in a laboratory setting as part of a program unrelated to therapy and treatment of behavior. The studies have shown that clients evince more severely problematic behavior in the laboratory setting. Thus, the set and setting contribute to the assessment results.

A related assessment problem pertains to the precise measures used to assess client behavior. The measures commonly used to eval-

uate therapy effects are paper-and-pencil inventories of specific psychological traits or states (e.g., anxiety), global self-evaluation, samples of direct behaviors, or physiological responses. Although these measures might well reflect change, they vary in the extent to which they reflect the problem for which the client may have sought treatment. As mentioned earlier, clients do not seek treatment because of the severity of a problem as reflected on *psychological measures*. Yet, it is these measures that are used to evaluate treatment. The extent to which a study contains measures that relate to the situations in which the client expresses his or her problem is important in generalizing the results of the study to the clinical situation.

The use of direct behavioral samples in contrived situations, if clients have complained of specific problematic behaviors, may approximate the clinical situation. For example, drinking of alcoholics has been assessed in inpatient treatment settings in which clients can sit and consume alcohol at a simulated bar (e.g., Sobell, Schaefer, & Mills, 1972). This might appear to be a close approximation of the clinical problem of drinking outside of the laboratory. However, it is still relatively distant from nonlaboratory situations. Indeed, reports suggest that alcoholics drink infrequently in treatment facilities when alcohol is made available (Skoloda, Alterman, Cornelison, & Gottheil, 1975). Apparently, many of the psychological cues of the natural environment (e.g., work demands, interactions with one's spouse) rather than physical cues of the drinking situation may precipitate drinking (Lawson, Wilson, Bridgell, & Ives, 1976). Thus, a method of assessment that reflects the behavior even more directly than a laboratory approximation would be more representative of clinical change. Alcohol consumption sampled at random periods in the client's natural environment would provide a more direct measure of the clinical behavior (e.g., Miller, Hersen, Eisler, & Watts, 1974). If alcohol consumption is assessed in the original situation in which the behavior is problematic, this is not an approximation of the clinical problem but a direct reflection of the problem itself. In any therapy investigation, it is important to

evaluate the extent to which the response measures approximate the client's problem and the manner in which assessment is conducted.

Degree of Resemblance to Clinical Situations

The above discussion outlines dimensions along which therapy investigations can vary. The list of dimensions is not necessarily exhaustive. The importance of the discussion does not derive from the specific dimensions enumerated. Rather, the value derives from viewing treatment investigations as multidimensional. The generality of the results of an experimental investigation of treatment depends on where the investigation lies with respect to these and similar dimensions vis-à-vis the clinical situation.

Each of the dimensions, and occasionally separate aspects of a given dimension, can be viewed as a continuum along which studies can vary. The continuum reflects the degree of resemblance of the study on a given dimension to the clinical therapy situation to which the results are to be generalized. The continuum for a given dimension ranges from identity with or close resemblance to the clinical situation to little or no resemblance to the clinical situation. In more commonly used terms, the continuum might be analogous to classifying the study as clinical (or applied) versus laboratory (or basic) research. However, this classification is an oversimplification, because it does not treat individually the different dimensions along which a study can vary and be evaluated.

If each of the dimensions mentioned earlier is viewed as a separate continuum along which the study might be evaluated, the task of evaluating an analogue study depends on where it falls on the continuum with respect to each dimension. Also, the extent to which the results of the study can be generalized depends on how the dimension relates to treatment efficacy. The issue here is how to decide the generality of the results of a study that in some way only resembles the clinical situation.

An explicit assumption in the therapy literature is that the degree to which a study resembles a clinical situation (for a given

dimension) indicates the likelihood of its generality to the clinical situation. Consider the dimension of severity of the target problem as one means of evaluating a therapy study. One would expect that a treatment shown to be effective in changing mildly problematic behavior (e.g., mild fear in college students) might have little generality to a more severe degree of similar behavior (clinical phobias). This is a very reasonable expectation and indeed might be bolstered with support from literature from both psychotherapy and medicine. For example, individuals who experience relatively severe insomnia do not respond as readily to placebo treatment as those who experience relatively mild or moderate insomnia (Nicolis & Silvestri, 1967). More generally, a concern in evaluating treatment is that less severe behaviors can be readily changed, whereas severe clinical problems cannot. This concern is the primary objection to analogue studies in behavior therapy in which the efficacy of select techniques occasionally has been based on the treatment of relatively mild problems (cf. Bernstein & Paul, 1971).

As a general rule, beyond considering only the target behavior dimension, an implicit assumption often made is that the greater the resemblance of a treatment investigation to the clinical situation, the more difficult it will be to change behavior. Changing behavior is considered to be increasingly difficult as the investigation departs from well-controlled laboratory analogue conditions and approaches characteristics of the clinical situation.

With many dimensions, it is possible and indeed likely that the assumed relationship is accurate. Thus, studies that only faintly resemble the clinical situation could readily produce changes that are not likely to be generalized to the clinical situation. However, with other dimensions, it is quite reasonable to expect that the results may be quite generalizable from laboratory to clinical settings. Possibly, the *less* resemblance of the study to the clinical situation for a given dimension, the *more* difficult it would be to change behavior. That is, departure from the clinical situation may make behavior change more difficult to achieve. In these cases, dem-

onstration of behavior change in the nonclinical situation might be a more convincing demonstration of an effect of treatment than would application in the clinical situation. For example, consider the *therapist* dimension. If therapists who bear very little resemblance to those who normally practice therapy (e.g., show less experience, education, and training) effect marked behavior change in their clients, the generality of this finding to highly trained therapists would be quite plausible. One might expect that if therapists who were actually untrained high school students, for example, could effectively change behavior, the effects would be likely to apply at least as well to trained professionals.

Similarly, consider the *treatment* dimension; if behavior change is shown with a variation of treatment that deviates from the clinical version, this does not necessarily mean that the results might not be generalized to the clinical version of the treatment. In many cases, the laboratory version of treatment is an important test of clinical treatment, because it minimizes the parameters of treatment that are likely to contribute to change (e.g., individualization of treatment, deviations from the treatment rationale to handle individual subject problems, etc.).

Finally, consider the *client* dimension. Clients who normally seek treatment in a clinical setting are likely to be experiencing some sort of crisis (i.e., defined as severity of the symptoms or behavioral problem) that precipitated seeking help. In this state, clients who seek help may be particularly motivated for and receptive to treatment and therefore hold high expectancies for improvement. This may make clients show improvements even before or very early in treatment (e.g., Frank, Nash, Stone, & Imber, 1963; Goldstein & Shipman, 1961). In addition, in the clinical situation, clients may be likely to comply with therapeutic instructions and perhaps even accept the treatment rationale relatively uncritically. Although all clients who seek treatment are not desperate for relief, severity of the problem should increase the incentive for attending and adhering to therapy.

At the other end of the continuum might be college students recruited for a treatment study with inducements of course credit. The

incentives for obtaining help and for complying with the requirement of treatment may be less than for clinical patients. If a given treatment effects change in subjects who are less motivated and who in fact adhere less rigorously to treatment requirements (e.g., do not perform "homework" assignments as conscientiously) than do clients in a clinical situation, then the effects of treatment could be more pronounced in a clinical situation.

The purpose of the present article is merely to raise the possibility rather than to assert that the analogue situation may occasionally provide a more rigid and more conservative test of a relationship between treatment and therapeutic change than that provided in the clinical situation. This is reasonable to entertain, because laboratory studies may dilute aspects of treatment that are central to behavior change. In these cases, the commonly assumed relationship between behavior change and the lack of resemblance of the study to the clinical situation is altered.

A problem in clinical research is that analogue research has been rejected by many on the grounds that by its very nature, it provides a weak test of the relationship between a variable and change in the clinical situation. Yet, this is not necessarily the case.

The relation between an analogue study and generality to clinical situations for a given dimension itself is an area of research. The importance of a given dimension (e.g., population, therapists, manner of recruitment, etc.) to the generality of the results needs to be evaluated directly. Increasing resemblance of a given dimension of a study to the clinical situation may not necessarily predict the extent to which the results apply to the clinical situation. In advance of studies showing this relationship, this assumption across all of the dimensions reviewed earlier should not be made.

The generality of results of a given study can only be evaluated directly by studying the variables of interest in the clinical situation itself. Thus, research always needs to test out the variables, insofar as possible, in the actual situation. However, the value of analogue research in understanding clinical phenomena can be partially evaluated

prior to direct clinical tests of individual findings. Research is needed that investigates the influence of departures from clinical situations along various dimensions and the implications of such departures for generalizing results to clinical situations. The effects of varying degrees of resemblance to the clinical situation for a given dimension need to be evaluated on therapy outcome. Comparisons of different points on a continuum varying in degrees of resemblance can reveal whether departure from the clinical situation for a given dimension attenuates, enhances, or has no effect on treatment. Once these relationships are elaborated, the generality of findings from research to clinical situations can be more easily evaluated for laboratory-based treatment studies.

Conclusion

Analogue therapy research has proliferated in recent years, especially in the area of behavior therapy. Extensive research has led to an empirically based methodology of carefully studying treatment procedures and parameters that influence behavior change. Yet, many authors have raised questions about the relevance of analogue work to the clinical situation and have suggested specific procedures to increase the similarity of treatment research to clinical applications (e.g., Bernstein & Paul, 1971; Borkovec & O'Brien, 1976). Research in psychotherapy and behavior therapy can differ from the clinical application of treatment along several dimensions such as the target problem, the clients and the manner in which they are recruited, the therapists, the selection of treatment, the client's set, and the setting in which treatment is conducted, the variations in treatment, and the assessment procedures used.

For evaluating the generality of research findings to the clinical situation, the main question of interest is not whether the investigation is an analogue of the clinical situation. An investigation invariably will be an analogue by the very nature of experimental research. *The alternative questions of interest seem to be the extent to which treatment research deviates from the clinical situation, along what dimensions, and whether the man-*

ner in which there are differences makes research a relatively strong or weak test of treatment in relation to the likely results in a clinical situation.

The primary purpose of the present article was to suggest a methodology for studying generality directly. Laboratory analogue research may differ from clinical treatment along several dimensions. Each dimension is a continuum spanning from little or no resemblance of the study to the clinical situation to close resemblance or identity with the clinical situation. Each investigation of treatment can be classified separately on several dimensions denoting its resemblance to the clinical situation. *Increasing similarity of an investigation to the clinical situation for a given dimension does not necessarily argue for greater generality of the results from the research to the clinical setting. Whether standing of a study on a given dimension is relevant to generality of the results needs to be determined empirically.* Laboratory treatment research that differs from clinical treatment may not necessarily provide a less stringent test of treatment. Indeed, for some dimensions, laboratory research may provide a more conservative test, and changes found in the laboratory may be even more likely to obtain in the clinical setting. However, the purpose here is not to speculate or make assumptions about the extent of generality of the findings of a study. It is hoped that dimensions that may affect generality will be subject to empirical scrutiny.

Reference Note

- 1 Bootzin, R. R. *Magnitude and durability of expectancy effects in behavior modification*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C., September 1971.

References

- Bergin, A. E. The evaluation of therapeutic outcomes. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis*. New York: Wiley, 1971.
- Bernstein, D. A. Behavioral fear assessment: Anxiety or artifact? In H. Adams & I. P. Unikel (Eds.), *Issues and trends in behavior therapy*. Springfield, Ill.: Charles C Thomas, 1973.
- Bernstein, D. A., & Nietzel, M. T. Procedural variation in behavioral avoidance tests. *Journal of Consulting and Clinical Psychology*, 1973, 41, 165-174.
- Bernstein, D. A., & Nietzel, M. T. Behavioral avoidance tests: The effects of demand characteristics and repeated measures of two types of subjects. *Behavior Therapy*, 1974, 5, 183-192.
- Bernstein, D. A., & Paul, G. L. Some comments on therapy analogue research with small animal "phobias." *Journal of Behavior Therapy and Experimental Psychiatry*, 1971, 2, 225-237.
- Borkovec, T. D., & O'Brien, G. T. Methodological and target behavior issues in analogue therapy outcome research. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 3). New York: Academic Press, 1976.
- Borkovec, T. D., Stone, N. M., O'Brien, G. T., & Kaloupek, D. G. Evaluation of a clinically relevant target behavior for analogue outcome research. *Behavior Therapy*, 1974, 5, 504-514.
- Borkovec, T. D., Wall, R. L., & Stone, N. M. False physiological feedback and the maintenance of speech anxiety. *Journal of Abnormal Psychology*, 1974, 83, 164-168.
- Brown, H. A. Role of expectancy manipulation in systematic desensitization. *Journal of Consulting and Clinical Psychology*, 1973, 41, 405-411.
- Cooper, A., Furst, J. B., & Bridger, W. H. A brief commentary on the usefulness of studying fears of snakes. *Journal of Abnormal Psychology*, 1969, 74, 413-414.
- Cowen, E. L. The experimental analogue: An approach to research in psychotherapy. *Psychological Reports*, 1961, 8, 9-10.
- Devine, D. A., & Fernald, P. S. Outcome effects on receiving a preferred, randomly assigned, or non-preferred therapy. *Journal of Consulting and Clinical Psychology*, 1973, 41, 104-107.
- Eysenck, H. J. The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, 1952, 16, 319-324.
- Frank, J. D. *Persuasion and healing: A comparative study of psychotherapy* (2nd ed.). Baltimore, Md.: Johns Hopkins University Press, 1973.
- Frank, J. D., Nash, E. H., Stone, A. R., & Imber, S. D. Immediate and long-term symptomatic course of psychiatric outpatients. *American Journal of Psychiatry*, 1963, 120, 429-439.
- Goldstein, A. P., & Dean, S. J. (Eds.). *The investigation of psychotherapy: Commentaries and readings*. New York: Wiley, 1966.
- Goldstein, A. P., & Shipman, W. G. Patient expectancies, symptom reduction and aspects of the initial psychotherapeutic interview. *Journal of Clinical Psychology*, 1961, 17, 129-133.
- Gordon, R. M. Effects of volunteering and responsibility on the perceived value and effectiveness of a clinical treatment. *Journal of Consulting and Clinical Psychology*, 1976, 44, 799-801.
- Hagen, R. L., Foreyt, J. P., & Durham, T. W. The dropout problem: Reducing attrition in obesity research. *Behavior Therapy*, 1976, 7, 463-471.

- Heller, K. Experimental analogues of psychotherapy: The clinical relevance of laboratory findings of social influence. *Journal of Nervous and Mental Disease*, 1963, 137, 420-426.
- Heller, K. Laboratory interview research as an analogue to treatment. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis*. New York: Wiley, 1971.
- Kazdin, A. E., & Wilson, G. T. *Evaluation of behavior therapy: Issues, evidence, and research strategies*. Cambridge, Mass.: Ballinger, 1978.
- Lana, R. E. Pretest sensitization. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press, 1969.
- Lang, P. J. Fear reduction and fear behavior: Problems in treating a construct. In J. M. Shlien (Ed.), *Research in psychotherapy* (Vol. 3). Washington, D.C.: American Psychological Association, 1968.
- Lawson, D. M., Wilson, G. T., Briddell, D. W., & Ives, C. C. Assessment and modification of alcoholics' drinking behavior in controlled laboratory settings: A cautionary note. *Addictive Behaviors*, 1976, 1, 299-303.
- Levis, D. J. The case for performing research on nonpatient populations with fears of small animals: A reply to Cooper, Furst, and Bridger. *Journal of Abnormal Psychology*, 1970, 76, 36-38.
- Luborsky, L., Singer, B., & Luborsky, L. Comparative studies of psychotherapies: Is it true that "everyone has won and all must have prizes"? *Archives of General Psychiatry*, 1975, 32, 995-1008.
- Meltzoff, J., & Kornreich, M. *Research in psychotherapy*. New York: Atherton, 1970.
- Miller, P. M., Hersen, M., Eisler, R. M., & Watts, J. G. Contingent reinforcement of lowered blood/alcohol levels in an outpatient chronic alcoholic. *Behaviour Research and Therapy*, 1974, 12, 261-263.
- Nicols, F. B., & Silvestri, L. C. Hypnotic activity of placebo in relation to severity of insomnia: A quantitative evaluation. *Clinical Pharmacology and Therapeutics*, 1967, 8, 841-848.
- Orlinsky, D. O., & Howard, K. I. *Varieties of psychotherapeutic experience*. New York: Teachers College Press, 1975.
- Orne, M. T. Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press, 1969.
- Paul, G. L. *Insight versus desensitization in psychotherapy: An experiment in anxiety reduction*. Stanford, Calif.: Stanford University Press, 1966.
- Paul, G. Behavior modification research: Design and tactics. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- Pope, K. S., Geller, J. D., & Wilkinson, L. Fee assessment and outpatient psychotherapy. *Journal of Consulting and Clinical Psychology*, 1975, 43, 835-841.
- Rachman, S. R. *The effects of psychotherapy*. Oxford, England: Pergamon Press, 1971.
- Singerman, K., Borkovec, T. D., & Baron, R. S. Failure of a "misattribution therapy" manipulation with a clinically relevant target behavior. *Behavior Therapy*, 1976, 7, 306-313.
- Skoloda, T. E., Alterman, A. I., Cornelison, F. S., & Gotthel, E. Treatment outcome in a drinking decision program. *Journal of Studies on Alcohol*, 1975, 36, 365-380.
- Sloane, R. B., Staples, F. R., Cristol, A. H., Yorkston, N. J., & Whipple, K. *Psychotherapy versus behavior therapy*. Cambridge, Mass.: Harvard University Press, 1975.
- Sobell, M. B., Schaefer, H. H., & Mills, K. C. Differences in baseline drinking behavior between alcoholics and normal drinkers. *Behaviour Research and Therapy*, 1972, 10, 257-267.
- Stuart, R. B. Notes on the ethics of behavior research and intervention. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice*. Champaign, Ill.: Research Press, 1973.
- Truax, C. B., & Mitchell, K. M. Research on certain therapist interpersonal skills in relation to process and outcome. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis*. New York: Wiley, 1971.
- Weber, S. J., & Cook, T. D. Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, 1972, 77, 273-295.
- Wilson, G. T. Effects of false feedback on avoidance behavior: "Cognitive" desensitization revisited. *Journal of Personality and Social Psychology*, 1973, 28, 115-122.
- Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford, Calif.: Stanford University Press, 1958.

Received April 25, 1977 ■

Methodological Considerations in Treatment Outcome Research on Obesity

G. Terence Wilson
Rutgers—The State University

Although there has been a dramatic increase in treatment research on obesity, numerous methodological inadequacies continue to obstruct further development of the field. Specific problems are identified, and corrective procedures are outlined. The importance of multiple measurement of outcome, including appropriate assessment of weight, activity, and relevant physical, emotional, and social indices of adjustment, is emphasized. Traditional evaluation criteria need to be broadened to encompass factors such as cost effectiveness of treatment. The relative merits of alternative single-subject and between-groups experimental designs are discussed, the latter including the treatment package, constructive, dismantling, and comparative research strategies. Ways of controlling for non-specific treatment influences are reviewed, and the contribution of client and therapist variables to outcome are mentioned. The reasons for conspicuous lack of long-term follow-ups are analyzed, and recommendations for reducing subject attrition during treatment and follow-up are advanced.

It was only 10 years ago that Stuart (1967) reported his findings on the behavioral treatment of overeating. Over the years that have followed this landmark publication, we have witnessed a dramatic upsurge of interest and research activity in the behavioral treatment of obesity. Research reports, evaluative reviews, books, and audiotape cassettes on the treatment of obesity seem to be appearing at an ever-quicken pace. Less evident but nonetheless revealing of this trend are the numerous manuscripts that are increasingly submitted to the psychological journals but that are rejected on the basis of one or other methodological problem. In short, as Jeffery, Wing, and Stunkard (1978) observed, behavioral treatment of obesity has attained a popularity among professionals that verges on faddism.

These developments in the treatment of obesity are attributable to several factors. In part, they reflect a more general phenome-

non, namely the extraordinary growth of behavior research and therapy as a whole (cf. Franks & Wilson, 1973, 1977; Hoon & Lindsley, 1974). The fact that obesity has been a subject of intensified research interest within the wider context of the development of behavior therapy is not surprising. Unlike most other clinical problems that have been the focus of treatment outcome research, obesity offers a definitive yet convenient and objective measure of outcome efficacy—weight reduction. Instead of the all-too-familiar difficulties inherent in the measurement of other clinical disorders, such as rating anxiety, inferring depression, judging interpersonal adjustment, or estimating alcohol consumption, a quantitative measure is obtained by simply asking the obese client to step on a scale. Furthermore, unlike most other behavior disorders, large samples of obese subjects for research investigations are readily available. Finally, largely as a result of historical accident—Stuart's (1967) influential application of the principles and procedures initially proposed by Ferster, Nurnberger, and Levitt (1962)—weight reduction has become an important arena for the eval-

Requests for reprints should be sent to G. Terence Wilson, Graduate School of Applied and Professional Psychology, Rutgers—The State University, P.O. Box 819, Piscataway, New Jersey 08854.

uation of behavioral self-control methods. The happy combination of a handy, "hard" index of treatment outcome for large numbers of motivated subjects who can usually be recruited at minimal effort and cost and who provide a made-to-measure testing ground for one of the hottest areas of behavioral modification has proven irresistible to behavioral researchers.

These circumstances have encouraged behavioral research on obesity. However, the ostensible ease with which controlled research can be conducted with obese clients has also helped spawn countless studies that have contributed neither to our knowledge about the treatment of obesity in particular nor to the development of the principles of behavior change in general. Many of these studies have been master's degree theses or doctoral dissertations in psychology. In contrast to those well-designed, thoughtful, and often time-consuming and effortful doctoral research projects that have advanced our knowledge, a large number of studies are reeled off more because of the convenience than for any other reason. Indeed, on the basis of several years of reviewing manuscripts for this and other psychological journals, I argue that short-term "quickie" studies on the behavioral treatment of obesity are now almost on a par with so-called "analogue" studies of snake-fearful college students as part of the staple diet of investigators seeking mainly to satisfy degree requirements and/or bolster sagging publication records. This practice, together with other largely unacknowledged influences discussed more fully below, have contributed directly to two of the major shortcomings in the field, namely, the paucity of long-term follow-up evaluations of treatment programs and the predominance of treatment findings that are frequently statistically but not clinically or substantively significant.

The present article is an attempt to describe some of the more common methodological inadequacies in research on the treatment of obesity and to suggest alternative, improved strategies for future evaluations. Research on the "new obesity therapy" has been accompanied by renewed investigation of presumed differences in eating patterns in obese

and nonobese populations; however, this article focuses on treatment outcome considerations. Although the issues raised in the following discussion have broad relevance to the evaluation of all treatment outcome research on obesity, the particular focus is on behavioral approaches.

Measurement of Treatment Outcome

Weight Loss

As stated above, the availability of an objective measure of weight loss was a *raison d'être* for the application and evaluation of behavioral principles in the modification of obesity. However, some studies, mostly those of a nonbehavioral nature, still fail to report specific information on weight loss, the custom being to report general impressions of changes in subjects' personal adjustment and emotional well-being (cf. Leon, 1976). The importance of gathering multiple measures of subjects' response to treatment is emphasized below. Yet, regardless of whatever other outcome measures are obtained, direct assessment of weight loss is a *sine qua non* of adequate evaluation of treatment of obesity. Neglecting to report precise data on weight in favor of a discussion of inferred psychological changes obfuscates evaluation of treatment efficacy and, as Bandura (1969) cautioned, can easily result in the "perpetuation of weak methods on the basis of extraneous criteria" (p. 540).

Although it is necessary that specific weight loss be reported, it is not sufficient. Clearly a 40-pound (18 kg) weight loss in a 400-pound (181.6 kg) individual is not the same as a 40-pound weight loss in a 160-pound (72.7 kg) individual. Following Stunkard and McLaren-Hume (1959), several studies have reported results in terms of percentage of subjects who lose specified amounts of weight, for example, more than 10 (4.5 kg), 20 (9.0 kg), 30 (13.5 kg), or 40 pounds, respectively. It is difficult to interpret these figures in the absence of precise information on subjects' initial weights. Musante (1976), for example, reported treatment outcome according to clients who lost more than 20 pounds or 40 pounds, respectively. Initial weights, however, ranged as widely as 155–484 pounds (70.5–

197 kg) in males and 125–321 pounds (56.8–145.9 kg) in females. As others have noted (cf. Hall & Hall, 1974; Jeffrey, 1975), simply reporting absolute weight loss severely impedes meaningful comparisons among different studies. Comparative analyses of the outcome of different studies requires the use of a standardized measure of improvement. Change in percentage overweight is more informative than absolute weight loss, since it takes initial weight into account. Better still is the weight reduction quotient (cf. Feinstein, 1959), which measures the amount of weight loss relative to the amount needed to obtain an ideal target weight. Specifically, the weight reduction quotient is a function of Relative Initial Overweight \times Percent of Surplus Weight Loss:

$$\text{weight reduction index} = \frac{\text{pounds lost}}{\text{pounds overweight}} \times \frac{\text{initial weight}}{\text{ideal weight}} \times 100.$$

This index takes account of height, amount overweight,¹ weight reduction goals, and absolute pounds lost. Several studies have used a truncated version of this index, namely,

$$\frac{\text{pounds lost}}{\text{pounds over ideal weight}} \times 100,$$

which reflects percentage of weight goal attained (e.g., Ashby & Wilson, 1977; Brownell, in press; Kingsley & Wilson, 1977; Mahoney, 1974). The complete formula compensates for the fact that the obese person has to lose more absolute weight to attain the same percentage of his or her target weight goal.

Both measures of weight—absolute weight loss and the reduction quotient—should be reported in treatment outcome studies. These two measures have yielded essentially similar patterns of results in several studies to date (e.g., Ashby & Wilson, 1977; Green, in press; Kingsley & Wilson, 1977; Mahoney, 1974). However, Brownell (in press) found a discrepancy between the two measures. Differences among treatment groups that were statistically significant in terms of absolute weight loss and change in percentage overweight were not statistically significant in terms of percentage of target weight goal attained.

Weight measurements should be based on specific readings obtained by the investigator using a reliable scale. Until the validity of self-reports of weight loss is established, individuals' estimates of their own weight gathered over the telephone (e.g., Hanson, Borden, Hall, & Hall, 1976) or through the mail (e.g., Mahoney & Mahoney, 1976) are less desirable and must be viewed with appropriate caution.

Treatment outcome measures based on weight have been criticized as true measures of obesity, that is, excessive body fat (LeBow, 1977). Overall weight reflects the person's total body mass, which includes the musculature and water composition in addition to body fat. Body weight might be a good predictor of body fat in very overweight individuals, but the relationship is less clear for minimal or moderate degree of overweight. Body weight is an especially poor index of body fat in young children, the elderly, and when comparisons are made among different age groups (Mahoney, Rogers, Straw, & Mahoney, in press). Since it is body fat that is presumably the target of obesity treatment programs, indices other than weight that specifically measures body fat have been advocated. Changes in body weight might confound reductions in body fat with the effects (usually temporary, however) of diuretics and laxatives. The use of skin-fold measures may provide a better predictor of body fat provided that precautions are taken to ensure reliability (Franzini & Grimes, 1976; Johnson & Stalonas, 1977) as a reliable and valid index of body fat. For these reasons, future outcome research might profitably include specific measures of body fat in addition to changes in body weight.

Objective Measures of Eating Behavior

There are instances in which the direct measurement of specific eating behaviors of food

¹ Ideal weight is usually based on the 1959 Metropolitan Life Insurance Company norms (U.S. Department of Health, Education and Welfare, 1967). Less frequently used are actuarial data of height and weight by age in a U.S. Department of Agriculture report (Hathaway & Foard, 1960). Seltzer (1965) discusses the limitations of such height-weight tables.

consumption patterns rather than changes in weight is of primary concern. Some examples and measurement strategies can be briefly noted. For example, Diamant and Wilson (1975) evaluated the efficacy of a particular behavioral treatment technique, covert sensitization, in decreasing consumption of a specific target food. Weight loss was too molar a dependent variable for this purpose, since it is a function of several variables that are not specifically affected by covert sensitization. Thus covert sensitization might decrease an individual's consumption of a particular food but not ensue in weight loss, because the individual increases consumption of other foods that were not targeted in the treatment. The efficacy of that technique would be obscured. This would be unfortunate, because if a specific method could be shown to be effective in reducing specific consumption of a particular food, then it might be extended to decrease other excessive eating habits or included in a broader treatment program that has as its aim the modification of obesity as such. On the other hand, if the technique is shown to be ineffective, then it can be dropped from multifaceted treatment programs. Accordingly, Diamant and Wilson (1975) adapted the taste-rating task from Schachter, Goldman, and Gordon (1968) as a direct laboratory measure of the consumption of specific target and nontarget foodstuffs rather than rely solely on subjective ratings of specific food preferences as an outcome measure of the effects of covert sensitization. (Some of the methodological considerations governing the use of the taste-rating task as an index of treatment outcome were raised by Leon & Roth, 1977; see also Nathan & Briddell's 1977, discussion of the assessment of alcohol consumption.)

The taste-rating task is a laboratory measure of eating behavior under controlled conditions. Measurement of food consumption in the natural environment is achieved by unobtrusive behavioral observations. Epstein, Parker, McCoy, and McGee (1976) and Gaul, Craighead, and Mahoney (1975) provide illustrations and discussion of the use of systematic behavioral observation as a means of assessing eating under natural conditions.

Direct Indicators of Cardiovascular Health

Mahoney et al. (in press) have criticized the almost exclusive reliance of treatment outcome studies on body weight. They suggest that more attention be given to factors that are directly related to cardiovascular health, for example, blood pressure, serum lipids (particularly triglycerides), and activity. The importance of exercise components has often been deemphasized in weight reduction programs. In lieu of less practical procedures, activity level can be assessed and caloric expenditure can be estimated on the basis of subjects' self-monitoring.

Criteria For Evaluating Treatment Outcome In Obesity

In addition to the specific measurement considerations discussed above, there are broader issues that bear on the evaluation of treatment programs for obesity. These issues include both client-related and efficiency and cost-related criteria for outcome evaluation.²

Clinical Significance of Treatment Effects

As noted above, the fact that the evaluation of behavioral methods rather than the treatment of obesity per se has frequently been the primary purpose of experimental investigations has resulted in findings that are statistically but not necessarily clinically significant (cf. Stunkard & Mahoney, 1976). Yet, in clinical research a major criterion for evaluating therapy is the overall importance of the treatment-produced improvement. The clinical significance of amount of weight reduction can be considered to be of comparable importance to the statistical comparison of group differences. The magnitude of weight loss needs to be given much more attention in the future treatment literature.

Proportion of Clients Who Improve

Designs that emphasize statistical comparison of group differences average the amount of weight reduction across all subjects within

² See Kazdin and Wilson (1978) for a more detailed discussion of these broader criteria for the evaluating treatment outcome in general.

each treatment group. This tends to obscure individual differences, a consequence of particular importance in the treatment of obesity in which massive interindividual variability has been one of the most consistent findings in the literature (cf. Jeffery et al., 1978; Penick, Filion, Fox, & Stunkard, 1971). Two different treatment methods might result in significant improvement in a greater proportion of subjects while some subjects show an increase in weight. For example, Penick et al. (1971) found considerably greater variability in weight losses in their subjects treated with behavior modification than in subjects treated with traditional psychotherapy. The five most successful subjects, as well as the single least improved subject, were in the behavioral treatment. Since a particular treatment might be recommended on the basis of the proportion of treated clients who are likely to show some specified improvement rather than on the basis of a group average, reports of treatment studies should indicate the number of individuals who lose clinically significant amounts of weight. Rather than simple categories of absolute pounds lost (e.g., Stunkard & McLaren-Hume, 1959), however, an index of weight corrected for initial relative obesity would be preferable. More specifically, given the recurring interindividual variability in treatment outcome that has yet to be adequately explained, it is suggested that individual data be reported even in the context of between-group designs (e.g., Aragona, Casady, & Drabman, 1975).

Multiple Measures of Change

In addition to the primary measure of weight loss, treatment studies should include concurrent evaluation of subjects' physical, emotional, and social functioning (Coates, 1977). For example, a therapeutic method might decrease weight but result in adverse side effects. To take an extreme example, even if a surgical intervention such as the jejunoileal bypass operation were effective in decreasing weight, its numerous serious side effects might contraindicate its use, especially cases of moderate obesity (cf. Chlouverakis, 1975). Treatment approaches that might be equally effective in decreasing weight might

produce sufficiently different side effects so as to recommend one over the other.

In one of the rare comparative studies that has addressed this issue empirically, Öst and Götestam (1976) reported that a pharmacological treatment (fenfluramine) produced more frequent and more protracted negative side effects than a behavioral treatment program. Stunkard and Rush (1974) have described the occurrence of depression concomitant with weight loss in clients receiving traditional psychotherapy. The theory that obese adults who lose weight fall below a biologically dictated "set point," and that as a result they enter a state of chronic energy deficit, specifically predicts adverse emotional reactions such as depression and irritability as a function of continued weight loss (cf. Nisbett, 1972). Yet the results of behavioral treatment programs have indicated no such effects. On the contrary, weight loss in behavioral programs has been associated with positive consequences for emotional and social adjustment (e.g., Stuart, 1967, 1971; Wollersheim, 1970). Adverse side effects of behavioral treatment techniques that produced weight loss have been noted. Contingency contracting resulted in significant weight loss, but closer analysis revealed that subjects adhered to the externally imposed reinforcement system by resorting to a variety of potentially hazardous procedures such as the use of diuretics, vomiting, starvation, and steam baths (cf. Mann, 1972). Contingency contracting has also been reported to increase absences from therapy sessions (Jeffrey, 1974) and other undesirable behavior (Mahoney, Moura, & Wade, 1973).

On the other hand, clients who do not lose significant amounts of weight might nonetheless show improvement in other areas of life functioning. Specific subjective and objective measures of the breadth and nature of the changes associated with the treatment for obesity will shed light on these potentially important relationships among different response systems.

Efficiency and Cost Effectiveness

Magnitude of weight reduction is not the only criterion for evaluating treatments for

obesity. The efficiency of treatment, including its duration, ease of administration, and disseminability, needs to be taken into account (cf. Kazdin & Wilson, 1978). Determining the cost effectiveness of treatment is particularly important given the availability of alternative treatment approaches of roughly comparable efficacy in producing weight reduction. Jeffrey (1975) has proposed a useful cost effectiveness index that is a function of the mean weight reduction quotient divided by mean treatment time per client. The financial cost of treatment is influenced by the costs of the professional training of the individuals who conduct treatment and the disseminability of the methods. In addition, the emotional cost experienced by the client in participating in a treatment program might be assessed. For example, procedures such as starvation diets and surgical interventions can be contrasted with behavioral methods that emphasize gradual weight loss over an extended time period by reducing overall caloric intake without prescribing any fixed diet. Consumer (client) evaluation of treatment would provide useful information about the acceptability of different methods. Some treatments might be inherently objectionable irrespective of financial cost or even outcome efficacy.

Experimental Design and Research Strategies

Treatment outcome research on obesity has consisted predominantly of between-group designs of one sort or another, although single-subject methodology has also been used. Each of these two contrasting research strategies includes several alternative methods replete with different methodological advantages and disadvantages.

Single-Case Experimental Methodology

Single-case experimental designs focus on therapeutic change in the individual client rather than average change across groups of clients. The defining characteristics of the designs include the observation of overt behavior, continuous assessment of change, and specific criteria for evaluating the reliability and significance of treatment-produced change

(see Hartmann & Hall, 1976; Hersen & Barlow, 1976; Kazdin, 1973).

There are several reasons why single-subject designs seem suited to research on the modification of obesity. First, the emphasis on the individual subject seems particularly appropriate given the repeated finding, noted above, of considerable interindividual variability in treatment outcome studies on obesity. Second, weight loss provides an observable and easily measured target behavior. And finally, single-subject methodology contains a therapeutic criterion for evaluating the clinical importance of behavior change. Specifically, this criterion emphasizes the magnitude of weight loss and the extent to which treatment brings the individual within acceptable or normative levels of body weight. This characteristic assumes special importance given that a major criticism of treatment studies has been the failure to demonstrate clinically meaningful reductions in weight as opposed to statistically significant comparisons among different treatment groups.

Some examples of the surprisingly few single-case experimental designs in the treatment of obesity can be briefly noted. Mann (1972) used an ABAB reversal design in demonstrating that contingency contracting can result in substantial weight losses. Weight loss occurred only when the contingency contracting procedure was in effect. Subjects lost no weight or even gained weight when it was withdrawn. Multiple-baseline designs have also been used. Morganstern (1974) used a multiple-baseline-across-behaviors design in showing that a self-administered aversion procedure decreased consumption of different target foods only when specifically applied to each food in sequence. Similarly, Epstein et al. (1976) described the use of a multiple-baseline-across-subjects design in demonstrating the role of instructions in the regulation of eating in obese and nonobese children.

A third form of single-case methodology, as yet to be used systematically in the modification of obesity, is the changing criterion design (Hartmann & Hall, 1976). In this design, following a stable baseline period, the treatment method is applied in order to reduce weight to a designated level. When this target weight is reached, the criterion (target

weight) is lowered further. The criterion is progressively lowered until the ultimate goal of weight reduction is attained. A functional relationship between treatment and weight change is said to be demonstrated if behavior (weight loss) repeatedly matches the criterion as the criterion is changed. Kazdin (1975) has pointed out that this design is particularly well-suited to behaviors that have to be shaped to reach a final goal. Emphasizing as it does the slow but steady reduction of weight, the behavioral treatment of obesity clearly fits this description.

Single-case experimental designs are not without their limitations (see Bandura, 1976; Franks & Wilson, 1977; Hersen & Barlow, 1976). In general, single-subject demonstrations offer no information of the generality of the findings to other cases. Nor do they permit evaluations of the comparative efficacy of different treatment methods. There are also specific problems. Reversal designs are uninterpretable if the pretreatment baseline does not recover when the therapeutic technique is withdrawn. A successful reversal design demonstrates a causal relationship between treatment technique and target behavior but, by definition, excludes one investigation of maintenance of treatment-produced change. The latter is of major importance in obesity research (see below). Interpretive problems are encountered with multiple-baseline designs when behavior change generalizes across behaviors, settings, or subjects. The changing criterion design cannot rule out confounding sources of behavior change even if weight losses consistently match the changing criterion. The introduction of each treatment phase can always be associated with events extraneous to the specific treatment method.

However, weight reduction as a target behavior appears to be relatively free of some difficulties that often beset single-subject methodology in applied settings. For example, it is usually possible to obtain stable baselines of sufficient duration before intervening with treatment, a fundamental requirement of an interpretable single-case experimental design. Clear-cut treatment effects that are discriminably different from pretreatment (baseline) level are commonplace. Furthermore, unlike many other target behaviors, normative data

exist that allow judgments about what constitutes a clinically significant weight loss.

Between-Groups Design

The methodological considerations that govern the use of between-groups designs are described in detail elsewhere (Campbell & Stanley, 1966; Paul, 1969). Suffice it here to address some specific issues that pertain directly to the evaluation of the treatment of obesity.

The treatment package strategy. Much of the outcome research on obesity has followed the treatment package strategy (cf. Kazdin & Wilson, 1978). In this strategy multiple techniques are administered as part of a single therapy program that is compared to a no-treatment or attention-placebo control condition. Assuming the internal validity of the study, a significant difference between the treatment and control groups can be taken to indicate a causal relationship. The purpose of this strategy is to achieve maximal weight reduction. Accordingly, as Azrin (1977) pointed out, these treatment package programs should "include as many component procedures as seem necessary to obtain, ideally, a total treatment success" (p. 144). Exemplifying many behavioral evaluations of behavioral programs, the Penick et al. (1971) study used a multi-component behavioral approach that included "everything but the kitchen sink" (Stunkard, 1976, p. 219). If this treatment package proves effective in producing weight loss, then experimental analyses of its various components can be undertaken to elucidate the reasons for treatment success and thereby subsequently refine and enhance the efficiency and efficacy of that treatment approach.

Among the problems to avoid in using this research strategy is the necessity of ensuring that the treatment package is not so complex and wide ranging that it becomes difficult to identify and distinguish among the specific techniques within such a multifaceted program. Moreover, the total program must be distinguishable operationally and conceptually from alternative treatment or control procedures against which it is compared. It follows that it is essential to describe treatment packages in a clear, operational fashion in order to

permit the reader of the manuscript to replicate the treatment procedure. All too often cursory reference is made to "the standard behavioral self-control program" or the "Stuart treatment program" in the description of procedures used to treat obesity. Vague, nonspecific descriptions such as these complicate evaluation and seriously hinder replication. Ideally, the investigator might make available the detailed treatment program manual to readers interested in further information about the procedure.

Constructive and dismantling research strategies. Broad-spectrum behavioral treatment programs have been consistently shown to produce greater weight loss than alternative approaches in the short term. Since a treatment effect has been demonstrated, component analyses of the active ingredients of these programs are appropriate targets of research. Experimental analyses of the efficacy of specific techniques, both singly and in combination, are especially called for in view of certain findings. Several studies have demonstrated that individual component parts of the broader behavioral treatment package can produce weight losses equal or superior to those effected with the full program (e.g., Bellack, 1976; Green, in press; McReynolds & Paulsen, 1976; Romanczyk, Tracey, Wilson, & Thorpe, 1973; see further discussion of this issue in Franks & Wilson, 1977).

There are two fundamental research strategies for analyzing the component parts of multifaceted treatment packages. In the constructive treatment strategy (McFall & Marston, 1970), the effect of a single, narrowly circumscribed treatment component is established, and extra components are added sequentially to determine if they enhance treatment effects. The effective components are retained as the larger treatment program is constructed. Illustrating the constructive strategy, Romanczyk et al. (1973) compared the following treatment groups: (a) no treatment control (waiting list); (b) self-monitoring control (weight only); (c) self-monitoring control (weight and calorie intake); (d) self-monitoring and symbolic aversion; (e) self-monitoring, symbolic aversion, and relaxation; (f) self-monitoring, symbolic aversion, and relaxation and behavioral management (stimu-

lus control) instructions; and (g) self-monitoring, symbolic aversion, relaxation, behavioral management, and contingency contracting. This design enabled the investigators to assess the incremental value of including different treatment techniques in the overall treatment package.

In the dismantling strategy (Lang, 1969), specific components are systematically eliminated and the associated decrement in treatment effect is measured. The relative contribution of each component to the total treatment package can then be assessed.

Comparative treatment strategy. A detailed discussion of the conceptual and methodological issues involved in a comparative treatment outcome research is beyond the scope of the present article (but see Kazdin & Wilson, 1978). Only some of the more pressing concerns are mentioned here. Well-controlled comparative outcome research entails the comparison of clearly specified, operationally distinguishable methods. Vague descriptions of therapeutic interventions in global terms such as "the Stuart behavioral treatment program," "group psychotherapy," or "insight therapy" are unacceptable. The specific procedures embraced by these general labels need to be spelled out. Similar problems attach to the comparison of treatment methods with something termed *routine treatment*. As a standard for comparison, treatments labeled as routine vary enormously in nature, ranging from those that include several active treatment ingredients to those that are minimal control conditions.

Of major importance is the need to ensure that an evaluation of any given treatment approach represents an adequate test of that approach. Behavioral methods are often compared with treatments described as "psychotherapy." In most instances the psychotherapy condition is more accurately construed as an attention-placebo control group. Drawing conclusions about the superiority of behavioral methods over traditional psychotherapy on the basis of these studies is inappropriate. The Penick et al. (1971) comparative outcome study provides a relatively rare example of the comparison of a behavioral program with an alternative method that was adequately representative of traditional therapy for obes-

ity. Öst and Götestam's (1976) study in which a behavioral method was contrasted with pharmacotherapy is another instance of a legitimate comparative outcome evaluation.

Different treatments must be distinguishable on an a priori basis and in their implementation. If these procedural differences are blurred, the results become uninterpretable. A study might be designed to evaluate the effect of a specific behavioral method such as self-monitoring of daily caloric intake. If, however, subjects in the comparison group not intended to engage in such self-monitoring learn about this procedure and implement self-monitoring themselves, the independent variable will have been compromised. The probable outcome of no differences between groups will be uninformative. Bellack, Rozensky, and Schwartz (1974) compared two forms of self-monitoring in which subjects monitored their behavior either before or after they ate. The two groups were not significantly different from each other. However, it is unclear whether subjects in the premonitoring condition complied with this procedure. Green (in press) had external observers systematically rate obese subjects' adherence to treatment instructions and found that fewer subjects in a premonitoring condition adhered to instructions than subjects in a postmonitoring condition. Green's finding that this temporal factor in self-monitoring did not significantly influence weight loss is therefore more interpretable than Bellack et al.'s data. Treatment procedures cannot be properly evaluated unless they are implemented, a factor that has possibly contributed to the great variability in treatment outcome with behavioral techniques.

The identification of the effective components of obesity treatment programs is impaired by the failure to assess subjects' adherence to treatment methods independently of outcome. The intent of most behavioral programs, for example, has been to alter eating habits as a means of producing weight loss. Typically, the independent variables of studies (manipulation of eating habits) have been inferred, quite inappropriately, from treatment outcome. Several strategies exist for assessing subjects' adherence to the treatment program. Both Hagen (1974) and Wollersheim (1970) used an eating patterns ques-

tionnaire to assess habit change. They reported significant positive correlations between weight loss and questionnaire responses. However, a questionnaire measure administered only once at the end of treatment is probably the least satisfactory means of determining subjects' eating patterns. Using subjects' self-monitoring records of daily eating behavior during the 1st and 5th weeks of treatment, Jeffery et al. (1978) found that changes in eating patterns were unrelated to weight loss. Similarly, Bellack et al. (1974) found no relationship between subjects' self-ratings of adherence to treatment guidelines and weight loss. Brownell (in press) used a similar but more comprehensive approach by assessing subjects' daily self-reports of 38 different weight-related behaviors over the entire treatment phase. Correlations among these measures and weight loss were nonsignificant. Continuous assessment offers obvious advantages.

Brownell (in press) had subjects' spouses rate their eating patterns as a means of obtaining estimates of their accuracy. Similarly, both Green (in press) and Öst and Götestam (1976) trained observers to rate subjects' target behavior. The latter reported a significant correlation between habit change and weight loss. The use of external observers provides a reliability check of subjects' self-reports of behavior change. However, the investigator should be alert to the possibility that this procedure is reactive. Green, for example, found that the use of external observers was differentially reactive across different treatment methods. (See Nelson, 1977, for a discussion of the reliability and reactivity of self-monitoring.)

Control Groups for Treatment Evaluation

The function of control groups is to ensure the internal validity of the study by precluding the effects of weight change over time that are independent of treatment and the nonspecific influences of the treatment itself. As a rule the nature of the control group that is included will depend on the specific research strategy and the precise question that is the focus of investigation. However, some general comments are in order.

The most basic control group is the no-treatment condition. There are now sufficient data to show that weight does not change significantly as a function of the mere passage of time, the effects of repeated assessments, and other factors that the no-treatment condition is designed to control for. It is no longer a necessary control in a between-group study, and it is definitely not sufficient. Simply comparing a treatment method with a no-treatment control group is no longer acceptable. Nonspecific³ treatment control groups are necessary if causal relationships between specific therapeutic techniques and weight loss are to be demonstrated. Nonspecific influences can be controlled in one of two ways. In one, the study might include at least two treatment groups that differ from each other, as in the dismantling or comparative strategies, for example, but that incorporate the nonspecific influences of the therapeutic process. A difference between the groups would be attributed to a specific treatment effect rather than a nonspecific influence. Bellack (1976), for example, showed that a self-reinforcement method that included self-monitoring was more effective than self-monitoring alone in producing weight loss, a difference that was ascribed to self-reinforcement, since nonspecific factors were presumably equated across conditions.

The other means of controlling for nonspecific influences involves the inclusion of a pseudotherapy or attention-placebo control group, exemplified by Wollersheim's (1970) pioneering study. Contrary to the view that it has been shown to be relatively unimportant (Jeffrey, 1974), this form of control constitutes a necessary feature of a well-designed outcome study. One reason is that more recent research has shown that it is more difficult to control for nonspecific influences than was once supposed. Many attention-placebo control treatments have been found to be less credible to subjects than the active treatment (cf. Borkovec & Nau, 1972; Kazdin & Wilcoxon, 1976). In other words, independent assessment has indicated that not all attention-placebo control conditions are successful in equating nonspecific treatment influences, such as expectations of therapeutic improve-

ment. These findings are relevant to research on obesity. For example, relaxation training has been used as a control condition for nonspecific effects in some studies (cf. Hall, Hall, Hanson, & Borden, 1974; Hanson et al., 1976). Yet Ashby and Wilson (1977) found that obese subjects consistently rated the relaxation training component of their multifaceted behavioral program as being of little use. It is not unreasonable to infer that its use as a pseudotherapy control condition would be less credible than treatment involving other behavioral techniques. Studies should include independent evaluations of the efficacy of attention-placebo control conditions in controlling for nonspecific influences (e.g., Kingsley & Wilson, 1977).

Another reason for retaining nonspecific control groups in treatment outcome studies is the fact that they have resulted in significant weight loss in some cases. Kingsley and Wilson (1977), for example, demonstrated that a social pressure control group modeled after one of Wollersheim's (1970) control groups (which she had found to have little effect on weight) was effective in producing weight loss. The effects of this control condition were especially evident at long-term follow-up, an evaluation few outcome studies ever make. The state of the art is far too undeveloped to forego the necessity of incorporating stringent and verifiable nonspecific control groups in treatment outcome studies on obesity.

Simple inclusion of an attention-placebo control group is insufficient to control for nonspecific influences in certain situations. Specifically, it does not allow for evaluation of a treatment method unconfounded by expectations of behavior change and the general demand characteristics of the experimental or therapeutic setting. One solution to this problem is the use of countertherapeutic (Diament

³ The term *nonspecific* is misleading. The different treatment influences indiscriminately lumped together under this accommodating rubric are quite specific. It is more realistic to propose that although nonspecific factors remain to be specified, they are neither intrinsically unspecifiable nor qualitatively different from other variables involved in planned behavior change (cf. Wilson & Evans, 1976).

& Wilson, 1975; Steinmark & Borkovec, 1974) or nondemand (Abrams & Wilson, Note 1) instructions. For example, Diamant and Wilson (1975) evaluated the efficacy of covert sensitization on a taste-rating laboratory eating task. Both the covert sensitization and attention-placebo control groups participated in two taste-rating tasks, which they were led to believe were simply baseline measures preparatory to the "real" treatment. Although both groups in effect received treatment during the period between these two assessments, countertherapeutic demands were communicated that emphasized that they would not lose weight until the real treatment began. The two behavioral measures served as pretreatment and posttreatment measures of food consumption.

Miscellaneous Methodological Issues

Description of Subjects

Appropriate background and demographic data on subjects should be reported. A major problem in the obesity treatment literature is the absence of any reliable predictors of successful treatment outcome. Post hoc correlations among weight loss and diverse life history and personality variables have been unrevealing (Bellack, 1975; Stunkard & Mahoney, 1976). In general, the more overweight the subject, the more weight loss that can be expected. Accordingly, Jeffery et al. (1978) caution that initial weight should be controlled for in the analysis of other predictor variables. Rather than focusing on personal characteristics of obese clients, the search for predictor variables might be more profitable if attention is directed to *how* they respond to the initial treatment program (Jeffery et al., 1978) and to performance on specific treatment program-related tasks (Bellack, 1975). As with personality assessment in general, the emphasis should be on what the subject *does* in relation to specific controlling variables rather than what the subject is like (Mischel, 1968). Finally, the manner in which subjects are recruited for treatment studies should also be spelled out.

Subject attrition. High dropout rates in obesity treatment studies are not uncommon

(Stunkard & McLaren-Hume, 1959). Careful reporting of the precise number of subjects in each group that drop out of treatment is essential. An attempt might be made to ascertain the reasons for attrition and to assess the progress of dropout subjects throughout the treatment period whenever possible. Data collected from these subjects might be included in statistical analyses. Interpretation of treatment findings must be guided by the fact that subjects who drop out of treatment are almost certainly treatment failures in obesity (cf. Levitz & Stunkard, 1974) and other (cf. Kent, 1976; Sobell & Sobell, 1976) treatment outcome studies.

Description of Therapists

In its emphasis on specific treatment techniques, therapy outcome research on obesity has tended to gloss over the potential contribution to weight loss of the therapists who administer the treatment program. Although the therapists in behavioral weight reduction studies have ranged from undergraduates with no therapeutic experience (e.g., Jeffery, 1974) to graduate students in clinical psychology (e.g., Romanczyk et al., 1973) to experienced professionals (e.g., Levitz & Stunkard, 1974), this factor is often overlooked. It is not uncommon for manuscripts to provide virtually no information about therapists. Yet a clear statement of the therapists' sex, their degree status, and their therapeutic experience in general and familiarity with the treatment of obesity in particular is imperative. The importance of this information is underscored by the finding that professionally trained therapists were significantly more effective than nonprofessionals (leaders of lay self-help groups) in reducing client attrition during therapy and in effecting weight loss (Levitz & Stunkard, 1974). Similarly, Jeffery et al. (1978) reported that therapists who were more experienced in the conduct of obesity treatment groups achieved results significantly superior to those of less experienced therapists. The influence of therapist variables on treatment outcome is not surprising, and has been observed in the application of even highly structured behavioral techniques to clinical

disorders other than obesity (cf. Alexander, Barton, Schiavo, & Parsons, 1976; Johnston, Lancashire, Mathews, Munby, Shaw, & Gelder, 1976; Wilson & Evans, 1976).

The potential influence of therapist variables can be controlled and evaluated statistically by using more than one therapist in a design in which each therapist administers comparable amounts of therapy to all groups. If a single therapist administers all treatment conditions, it is impossible to disentangle the relative contributions of therapist versus treatment method. Another source of Therapist \times Treatment confounding occurs if different therapists separately administer different treatment methods to different groups (e.g., Penick et al., 1971). In the latter study, experienced professionals conducted the traditional psychotherapy treatment, whereas therapists who had had no experience in behavior therapy, and little therapeutic experience in general, administered the behavioral method. This strategy explicitly "stacks the deck" against the treatment administered by the naïve therapists. To the extent that this particular treatment proves to be superior, irrespective of therapists' expertise, it provides a powerful demonstration of the efficacy of that method. However, if the superiority is not clear-cut, or if the treatment administered by the inexperienced therapists proves to be less effective, then the results cannot be interpreted unequivocally. Thus Penick et al. (1971) found that behavior therapy produced greater weight losses, although the differences were not consistently statistically significant. Especially in view of Levitz and Stunkard's (1974) findings, it is entirely plausible that had the behavioral method been administered by experienced therapists, its superiority would have been more pronounced and its greater variability in outcome reduced.

Other Procedural Details

Descriptions of treatment programs should include the therapeutic rationale and treatment instructions, the scheduling of sessions, policies regarding absences, and arrangements for any make-up sessions. The time of year during which a study is conducted should also

be stated in view of reports that subjects tend to lose more weight in the spring and summer than in the fall and winter (cf. Ashby & Wilson, 1977; Jeffery et al., 1978; Mahoney & Mahoney, 1976).

Follow-up Evaluation

A major shortcoming in research on the treatment of obesity is the relative lack of long-term follow-up evaluations of therapeutic efficacy (cf. Wilson, in press). The significance of this striking deficiency is highlighted by the fact that obesity is a clinical disorder that has been characterized by consistently high relapse rates; that is, clients who lose weight during treatment usually regain it (cf. Stunkard, 1976; Stunkard & Mahoney, 1976). Yet although the importance of long-term follow-ups is consistently stressed, the overwhelming majority of manuscripts that are submitted for publication nonetheless fail to include appropriate follow-ups. It is as important to enquire into the reasons for this discrepancy as it is imperative that future treatment outcome research include long-term follow-ups.

Long-term follow-ups require the investment of considerable time and effort. Whether they can be completed successfully is often uncertain and unpredictable. Consider then the fact that a great deal of the research on obesity is conducted by graduate students and faculty in primarily academic settings. Clearly, long-term research—given these characteristics—is not the ideal stuff of which doctoral dissertations are made of. Under the current system of doctoral dissertation research training, most students would not have the time to conduct a long-term follow-up themselves; the realities of the situation appear to reward instead more manageable, more predictable, and time-limited studies. The importance of treatment outcome (especially long-term follow-up), as Azrin (1977) has suggested, becomes secondary. The contingencies governing the research behavior of faculty members seeking promotion and tenure in the increasingly competitive job market are not dissimilar. Outcome research emphasizing long-term follow-up represents something of a gamble; it is not optimal for

boosting publication frequencies and enhancing curricula vitae. In sum, if we are to reverse these contingencies—and, after all, behavior is at least partly a function of its consequences—we must encourage dissertation research that is relevant to treatment outcome. A student's participation in one facet (e.g., the long-term follow-up) of a broader research program might be viewed as a legitimate dissertation topic. This can—indeed *must*—occur without relaxing rigorous standards of scholarship and methodological sophistication. Our interpretation of research productivity of faculty must extend beyond the number of completed studies and incorporate an emphasis on the nature of the research.

Outcome evaluation of psychological treatments for obesity should distinguish among the initial treatment-produced weight loss, its generalization to the natural environment, and its maintenance over time (Bandura, 1969). Different factors might govern each of these processes, and it is now clear that generalization and maintenance will be ensured only to the degree that specific strategies toward that end are used (e.g., Kingsley & Wilson, 1977). The methodological strictures for the investigation and evaluation of treatment apply with equal force to maintenance or the follow-up phase of the study. All procedures followed should be explicitly described. Weight measures should be taken directly rather than relying on subjects' self-reports over the telephone or via mail. If different maintenance strategies are being compared, it is essential to ensure that they are actually implemented and that they are conceptually and operationally distinct.

In some instances subjects seek additional or alternative sources of treatment during the follow-up period. This creates problems of interpretation (cf. Kazdin & Wilson, 1978) but cannot be avoided. It would be both futile and unethical to attempt to dissuade subjects from seeking extra assistance for their problem. However, the nature and extent of these other therapeutic experiences should be assessed and reported.

Attrition Rates

Dropouts are more common during follow-up than the initial treatment phase and just as damaging to the interpretation of results. External validity is jeopardized, since attrition reduces generalizability; internal validity is compromised when there is differential attrition across treatment groups. Accordingly, every effort must be made to minimize attrition rates. Several procedures for accomplishing this crucial goal may be mentioned briefly (but see more detailed discussions of this issue in Hagen, Foreyt, & Durham, 1976; Sobell & Sobell, 1976).

1. Maintenance procedures and long-term follow-up should be an integral part of the design of a treatment outcome study and should be presented to subjects as such. In committing themselves to the treatment program, subjects also explicitly commit themselves to the follow-up. It is *not* something to be tacked on following treatment almost as an afterthought.

2. Investigators might remain in frequent contact with subjects. Regular phone calls or contact via the mail are recommended. Related to frequency of contact is the nature of that contact. Personal contact that represents an extension of the initial treatment phase is indicated. Since this continuing contact could well be reactive, that is, function as a type of maintenance strategy in its own right (Sobell & Sobell, 1976), precise procedural description is necessary.

3. A refundable deposit that subjects forego if they miss too many sessions or dropout significantly reduces subject attrition (e.g., Hagen et al., 1976).

Christensen (1976) has argued that current follow-up methods might provide weight measures that are unrepresentative of subjects' eating habits. Specifically, subjects might resort to drastic short-term means (e.g., fasting or diuretics) of reducing weight immediately prior to a follow-up weigh-in evaluation (cf. Mann, 1972). Christensen suggested that fixed follow-up dates be replaced by randomly scheduled weigh-ins. However, some subjects might object to this method. In any event, frequent contact during follow-up that stresses the importance of the personal rela-

tionship between subject and investigator would provide a detailed, representative assessment of subjects' eating habits and weight patterns.

Reference Note

1. Abrams, D. B., & Wilson, G. T. *Self-monitoring, reactivity and smoking behavior*. Unpublished manuscript, Rutgers—The State University, 1977.

References

- Alexander, J. F., Barton, C., Schiavo, R. S., & Parsons, B. Systems-behavioral intervention with families of delinquents: Therapist characteristics, family behavior, and outcome. *Journal of Consulting and Clinical Psychology*, 1976, 44, 656-664.
- Aragona, J., Cassady, J., & Drabman, R. S. Treating overweight children through parental training and contingency contracting. *Journal of Applied Behavior Analysis*, 1975, 8, 269-278.
- Ashby, W. A., & Wilson, G. T. Behavior therapy for obesity: Booster sessions and long-term maintenance of weight loss. *Behaviour Research and Therapy*, 1977, 15, 451-464.
- Azrin, N. H. A strategy for applied research: Learning based but outcome oriented. *American Psychologist*, 1977, 32, 140-149.
- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Bandura, A. Self-reinforcement: Theoretical and methodological considerations. *Behaviorism*, 1976, 4, 135-155.
- Bellack, A. S. Behavior therapy for weight reduction. *Addictive Behaviors*, 1975, 1, 73-82.
- Bellack, A. S. A comparison of self-reinforcement and self-monitoring in a weight reduction program. *Behavior Therapy*, 1976, 7, 68-75.
- Bellack, A. S., Rozensky, R. H., & Schwartz, J. A comparison of two forms of self-monitoring in a behavioral weight reduction program. *Behavior Therapy*, 1974, 5, 523-530.
- Borkovec, T. D., & Nau, S. D. Credibility of analogue therapy rationales. *Journal of Behavior Therapy and Experimental Psychiatry*, 1972, 3, 257-260.
- Brownell, K. D. The effect of spouse training and partner cooperativeness in the behavioral treatment of obesity. *Behaviour Research and Therapy*, in press.
- Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1966.
- Chlouverakis, C. Dietary and medical treatments of obesity. An evaluative review. *Addictive Behaviors*, 1975, 1, 3-22.
- Christensen, A. Measuring and maintaining weight losses. *Behavior Therapy*, 1976, 7, 709-711.
- Coates, T. J. Theory, research, and practice in treating obesity: Are they really all the same? *Addictive Behaviors*, 1977, 2, 95-104.
- Diamant, C., & Wilson, G. T. An experimental investigation of the effects of covert sensitization in an analogue eating situation. *Behavior Therapy*, 1975, 6, 499-509.
- Epstein, L. H., Parker, L., McCoy, J. F., & McGee, G. Descriptive analysis of eating regulation in obese and nonobese children. *Journal of Applied Behavior Analysis*, 1976, 9, 407-415.
- Feinstein, A. R. The measurement of success in weight reduction: An analysis of methods and a new index. *Journal of Chronic Diseases*, 1959, 10, 439-456.
- Ferster, C. B., Nurnberger, J. I., & Levitt, E. B. The control of eating. *Journal of Mathematics*, 1962, 1, 87-109.
- Franks, C. M., & Wilson, G. T. *Annual review of behavior therapy: Theory and practice* (Vol. 1). New York: Brunner/Mazel, 1973.
- Franks, C. M., & Wilson, G. T. *Annual review of behavior therapy: Theory and practice* (Vol. 5). New York: Brunner/Mazel, 1977.
- Franzini, L. R., & Grimes, W. B. Skin-fold measures as the criterion of change in weight control studies. *Behavior Therapy*, 1976, 7, 256-260.
- Gaul, D. J., Craighead, W. E., & Mahoney, M. J. Relationship between eating rates and obesity. *Journal of Consulting and Clinical Psychology*, 1975, 43, 123-125.
- Green, L. The temporal and stimulus dimensions of self-monitoring in the behavioral treatment of obesity. *Behavior Therapy*, in press.
- Hagen, R. L. Group therapy versus bibliotherapy in weight reduction. *Behavior Therapy*, 1974, 5, 222-234.
- Hagen, R. L., Foreyt, J. P., & Durham, T. W. The dropout problem: Reducing attrition in obesity research. *Behavior Therapy*, 1976, 7, 463-471.
- Hall, S. M., & Hall, R. G. Outcome and methodological considerations in behavioral treatment of obesity. *Behavior Therapy*, 1974, 5, 352-364.
- Hall, S. M., Hall, R. G., Hanson, R. W., & Borden, B. L. Permanence of two self-managed treatments of overweight. *Journal of Consulting and Clinical Psychology*, 1974, 42, 781-786.
- Hanson, R. W., Borden, B. L., Hall, S. M., & Hall, R. G. Use of programmed instruction in teaching self-management skills to overweight adults. *Behavior Therapy*, 1976, 7, 366-373.
- Hartmann, D. P., & Hall, R. V. The changing criterion design. *Journal of Applied Behavior Analysis*, 1976, 9, 527-532.
- Hathaway, M. L., & Foad, E. O. *Heights and weights of adults in the United States*. (U.S. Department of Agriculture, Home Economics Research Report No. 10) Washington, D.C.: U.S. Government Printing Office, 1960.
- Hersen, M., & Barlow, D. H. *Single case experimental designs: Strategies for studying behavior change*. New York: Pergamon Press, 1976.
- Hoon, P., & Lindsley, O. R. A comparison of behavior and traditional therapy publication activity. *American Psychologist*, 1974, 29, 694-697.

- Jeffery, R. W., Wing, R. R., & Stunkard, A. J. Behavioral treatment of obesity: The state of the art. *Behavior Therapy*, 1978, 9, 189-199.
- Jeffrey, D. B. A comparison of the effects of external control and self-control on the modification and maintenance of weight. *Journal of Abnormal Psychology*, 1974, 83, 404-410.
- Jeffrey, D. B. Treatment evaluation issues in research on addictive behaviors. *Addictive Behaviors*, 1975, 1, 23-36.
- Johnson, W. G., & Stalonas, P. Measuring skinfold thickness—A cautionary note. *Addictive Behaviors*, 1977, 2, 105-108.
- Johnston, D. W., Lancashire, M., Mathews, A. M., Munby, M., Shaw, P. M., & Gelder, M. G. Imaginal flooding and exposure to real phobic situations: Changes during treatment. *British Journal of Psychiatry*, 1976, 129, 372-377.
- Kazdin, A. E. Methodological and assessment considerations in evaluating reinforcement programs in applied settings. *Journal of Applied Behavior Analysis*, 1973, 6, 517-531.
- Kazdin, A. E. *Behavior modification in applied settings*. Homewood, Ill.: Dorsey Press, 1975.
- Kazdin, A. E., & Wilcoxon, L. A. Systematic desensitization and nonspecific treatment effects: A methodological evaluation. *Psychological Bulletin*, 1976, 83, 729-758.
- Kazdin, A. E., & Wilson, G. T. *Evaluation of behavior therapy: Issues, evidence, and research strategies*. Cambridge, Mass.: Ballinger, 1978.
- Kent, R. A methodological critique of "Interventions for boys with conduct problems." *Journal of Consulting and Clinical Psychology*, 1976, 44, 297-298.
- Kingsley, R. G., & Wilson, G. T. Behavior therapy for obesity: A comparative investigation of long-term efficacy. *Journal of Consulting and Clinical Psychology*, 1977, 45, 288-298.
- Lang, P. J. The mechanics of desensitization and the laboratory study of fear. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- LeBow, M. Can lighter become thinner? *Addictive Behaviors*, 1977, 2, 87-94.
- Leon, G. R. Current directions in the treatment of obesity. *Psychological Bulletin*, 1976, 83, 557-578.
- Leon, C. R., & Roth, L. Obesity: Psychological causes, correlations and speculations. *Psychological Bulletin*, 1977, 84, 117-139.
- Levitz, L. S., & Stunkard, A. J. A therapeutic coalition for obesity: Behavior modification and patient self-help. *American Journal of Psychiatry*, 1974, 131, 423-427.
- Mahoney, M. J. Self-reward and self-monitoring techniques for weight control. *Behavior Therapy*, 1974, 5, 48-57.
- Mahoney, M. J., Moura, N. G. M., & Wade, T. C. Relative efficacy of self-reward, self-punishment, and self-monitoring techniques for weight loss. *Journal of Consulting and Clinical Psychology*, 1973, 40, 404-407.
- Mahoney, M. J., & Mahoney, K. Treatment of obesity: A clinical exploration. In G. J. Williams, S. Martin, & J. Foreyt (Eds.), *Obesity: Behavioral approaches to dietary management*. New York: Brunner/Mazel, 1976.
- Mahoney, B. K., Rogers, T., Straw, M. K., & Mahoney, M. J. *Human obesity: Assessment and treatment*. Englewood Cliffs, N.J.: Prentice-Hall, in press.
- Mann, R. A. The behavior-therapeutic use of contingency contracting to control an adult behavior problem: Weight control. *Journal of Applied Behavior Analysis*, 1972, 5, 99-109.
- McFall, R. M., & Marston, A. R. An experimental investigation of behavior rehearsal in assertive training. *Journal of Abnormal Psychology*, 1970, 76, 295-303.
- McReynolds, W. T., & Paulsen, B. K. Stimulus control as the behavioral basis of weight loss procedures. In G. J. Williams, S. Martin, & J. Foreyt (Eds.), *Obesity: Behavioral approaches to dietary management*. New York: Brunner/Mazel, 1976.
- Mischel, W. *Personality and assessment*. New York: Wiley, 1968.
- Morganstern, K. P. Cigarette smoke as a noxious stimulus in self-managed aversion therapy for compulsive eating: Techniques and illustrations. *Behavior Therapy*, 1974, 5, 255-260.
- Musante, G. J. The dietary rehabilitation clinic: Evaluative report of a behavioral and dietary treatment of obesity. *Behavior Therapy*, 1976, 7, 198-204.
- Nathan, P. E., & Briddell, D. W. Behavioral assessment and treatment of alcoholism. In B. Kissin & H. Begleiter (Eds.), *The biology of alcoholism* (Vol. 5). New York: Plenum Press, 1977.
- Nelson, R. O. Methodological issues in assessment via self-monitoring. In J. D. Cone & R. P. Hawkins (Eds.), *Behavioral assessment*. New York: Brunner/Mazel, 1977.
- Nisbett, R. Hunger, obesity, and the ventromedial hypothalamus. *Psychological Review*, 1972, 79, 433-470.
- Öst, L., & Götestam, K. Behavioral and pharmacological treatments for obesity: An experimental comparison. *Addictive Behaviors*, 1976, 1, 331-338.
- Paul, G. L. Behavior modification research: Design and tactics. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- Penick, S. B., Filion, R., Fox, S., & Stunkard, A. J. Behavior modification in the treatment of obesity. *Psychosomatic Medicine*, 1971, 33, 49-55.
- Romanczyk, R. G., Tracey, D. A., Wilson, G. T., & Thorpe, G. L. Behavioral techniques in the treatment of obesity: A comparative analysis. *Behaviour Research and Therapy*, 1973, 11, 629-640.
- Schachter, S., Goldman, R., & Gordon, A. Effects of fear, food deprivation, and obesity on eating. *Journal of Personality and Social Psychology*, 1968, 10, 91-97.

- Seltzer, C. C. Limitations of height-weight standards. *New England Journal of Medicine*, 1965, 272, 1132.
- Sobell, M. B., & Sobell, L. C. Second-year treatment outcome of alcoholics treated by individualized behavior therapy: Results. *Behaviour Research and Therapy*, 1976, 14, 195-216.
- Steinmark, S. W., & Borkovec, T. D. Active and placebo treatment effects on moderate insomnia under counterdemand and positive demand instructions. *Journal of Abnormal Psychology*, 1974, 83, 157-163.
- Stuart, R. B. Behavioral control of overeating. *Behaviour Research and Therapy*, 1967, 5, 357-365.
- Stuart, R. B. A three-dimensional program for the treatment of obesity. *Behaviour Research and Therapy*, 1971, 9, 177-186.
- Stunkard, A. J. *The pain of obesity*. Palo Alto, Calif.: Bull, 1976.
- Stunkard, A. J., & Mahoney, M. J. Behavioral treatment of the eating disorders. In H. Leitenberg (Ed.), *Handbook of behavior modification and behavior therapy*. New York: Appleton-Century-Crofts, 1976.
- Stunkard, A. J., & McLaren-Hume, M. The results of treatment for obesity. *Archives of Internal Medicine*, 1959, 103, 79-85.
- Stunkard, A. J., & Rush, J. Dieting and depression re-examined: A critical review of reports of untoward responses during weight reduction for obesity. *Annals of Internal Medicine*, 1974, 81, 526-533.
- United States Department of Health, Education, and Welfare. *Obesity and health: A source-book of current information for professional health personnel*. Arlington, Va.: U.S. Public Health Service, 1967.
- Wilson, G. T. Behavioral treatment of obesity: Maintenance strategies and long-term efficacy. In S. Bates, W. Docken, K. G. Götestam, L. Melin, & P. O. Sjöden (Eds.), *Trends in behavior therapy*. New York: Academic Press, in press.
- Wilson, G. T., & Evans, I. M. Adult behavior therapy and the therapist-client relationship. In C. M. Franks & G. T. Wilson (Eds.), *Annual review of behavior therapy: Theory and practice* (Vol. 4). New York: Brunner/Mazel, 1976.
- Wollersheim, J. P. Effectiveness of group therapy based upon learning principles in the treatment of overweight women. *Journal of Abnormal Psychology*, 1970, 76, 462-474.

Received October 19, 1977 ■

Smoking-Cessation Research

Richard M. McFall
University of Wisconsin—Madison

Methodological problems associated with treatment research on cigarette smoking are explored, and possible solutions are discussed. The main problems considered are the selection and retention of subjects, the measurement of smoking, the design of treatment studies, and the interpretation and generalizability of experimental results.

The Surgeon General's Report (U.S. Public Health Service, 1964) on the health hazards of tobacco stimulated a wave of smoking research that has persisted into the present with little sign of abating. In fact, there has been so much published theory and research on smoking that it has spawned a secondary publishing enterprise consisting of periodic literature reviews (e.g., Bernstein, 1969; Lichtenstein & Danaher, 1976), scholarly books and conference reports (e.g., Borgatta & Evans, 1968; Hunt, 1970), and a monthly bibliographic bulletin (*Smoking and Health Bulletin*, published by the National Clearinghouse for Smoking and Health, Atlanta, Georgia).

Psychological research on smoking generally falls into one of four categories: studies of the causes of smoking, studies of its effects, studies of its prevention, and studies of various treatment approaches. Thus far the greatest share of attention seems to have been devoted to the last category, that is, to the search for effective methods of helping cigarette smokers reduce or eliminate their smoking behavior.

Despite the continuing interest in smoking-cessation research, investigators have shown surprisingly little awareness of the special methodological problems inherent in such research, or in the ways to avoid or overcome such problems. Methodological articles (e.g., Bernstein, 1969) seem to have been ignored

by some investigators. The present article is an attempt to draw attention, once again, to the critical methodological issues. In preparing the article, I have drawn illustrative material both from published manuscripts and from manuscripts that were submitted for publication but were found unacceptable. Since the purpose of the article is to promote better future research, rather than to criticize previous investigations, the sources of most examples will not be referenced.

Subject Issues: Problems of Generalizability

Who are the subjects? Smoking research typically is not conducted with a randomly selected sample of subjects from the entire population of smokers.¹ Similarly, smoking-cessation studies typically do not draw their subjects randomly from among all persons who may wish to quit smoking. Rather, the subjects in smoking studies are nearly always volunteers or recruits, whose relation to the parent populations of smokers or aspiring quitters is virtually impossible to determine. This fact immediately limits the generalizations that can be drawn from the results of specific smoking studies. For example, when volunteer subjects are used to study the personality correlates of smoking, one cannot assume that the findings are relevant to all

Requests for reprints should be sent to Richard M. McFall, Department of Psychology, W. J. Brogden Psychology Building, University of Wisconsin, 1202 West Johnson Street, Madison, Wisconsin 53706.

¹ Of course, the use of actual smokers as subjects in smoking-cessation studies allows greater freedom to generalize about the results than if analogue subjects had been used, as is the case in some other areas of behavior modification research.

smokers; it is possible that the findings may be more a function of the subjects' willingness to volunteer than their smoking habits. Similarly, if a smoking-cessation program either succeeds or fails with a particular sample of volunteers, there is no guarantee that the same effect would be achieved with smokers who chose—for whatever reason—not to volunteer. In fact, one might argue that the volunteer subjects in smoking-cessation programs are not representative of smokers in general.

Perhaps the typical volunteer is a member of a subgroup of smokers who are seeking an externally imposed, somewhat magical, solution to their problem. Furthermore, as the more promising candidates from among this subgroup manage to quit smoking, the remaining subjects may represent a distilled, hard-core group of treatment-resistant smokers. Only a small percentage of the volunteer subjects in most smoking-cessation studies manage to quit smoking and to remain abstinent for at least 6 months. Seldom do 20% achieve this measure of success! Based on such results, some investigators have concluded, rather pessimistically, that smoking-cessation techniques generally are not effective. It is interesting to note, however, that in the meantime many individuals seem to have given up smoking on their own—without volunteering as subjects in formal treatment programs. Thus, over the course of numerous smoking-cessation studies, the characteristics of the volunteer subgroup may be changing.

As if the inherent limitations of working with nonrandom samples were not enough, some investigators further limit the generality of their results by failing to report in detail (a) how they recruited their subjects and (b) the essential characteristics of their resulting sample. It would be valuable if investigators reported whether volunteer subjects were recruited from introductory psychology classes, by newspaper ads, or through referrals from physicians; it also would help to know specifically what the subjects were told that enticed them to volunteer. At the very least, investigators should always provide descriptive statistics on their sample's age, sexual composition, occupational and educational status, living arrangements, cur-

rent smoking behavior, and smoking history. The smoking history should include information concerning the pattern and form of tobacco consumption over the years, the chronicity of the habit, the history of any prior attempts to quit, and any significant health problems.

Another important question about the subjects in treatment studies is, what are their personal goals? So-called volunteers in smoking programs actually may be responding to cultural, social, medical, or family pressures; these subjects may be very different from those who volunteer without such coercive pressures. Even among genuine volunteers, however, there may be important differences in personal goals: Some may be committed to achieving total abstinence, whereas others may be satisfied with a significant reduction in their smoking. Few smoking studies have asked subjects beforehand to indicate their reasons for volunteering or to state their personal treatment objective.

Subject mortality. The problem of "subject mortality" (see Campbell & Stanley, 1963) is one of the biggest methodological problems, or sources of invalidity, in smoking-cessation research. If subjects are randomly assigned to different treatment conditions but some subjects drop out of the experiment prematurely, this loss seriously undercuts the investigator's ability to interpret and generalize from the experimental results. There is no way to rule out the possibility that the subject loss has been nonrandom, thereby rendering the treatment groups no longer comparable. Asking subjects why they dropped out of the study is no remedy; even when subjects give reasons that seem unrelated to the experiment, these explanations may be little more than polite excuses or rationalizations. The fact that equal numbers of subjects may have dropped out of the different treatment groups does not mean that the mortality problem has been avoided; it may be that different kinds of subjects dropped out of the different groups, thus making them no longer comparable in composition, although they remain comparable in size. Replacing dropouts with new subjects does not solve the problem either; the replacement may be very different from the original sub-

ject, thus altering the group composition. And there simply are no acceptable post hoc methods for statistically correcting for subject mortality by artificially matching or equating groups. The only valid solution is to retain all original subjects.

Unfortunately, volunteers for smoking-cessation programs are notoriously unfaithful subjects. Without using some kind of special inducement to stay in a study or constraint against dropping out, it has been virtually impossible to retain an adequate proportion of the original sample in most smoking studies.

One fairly effective method of inducing subject fidelity has been to require an "earnest deposit" before admitting a subject to a smoking-cessation program. For example, some experimenters have collected \$20 or \$25 from each subject at the time of admission and have promised to return the full amount at the end of the study, contingent on the subject's faithful participation. The deposit money was to be refunded regardless of the subject's treatment outcome, so long as the subject attended treatment sessions and provided the requested smoking data. Other investigators have used variations of this method. In one variation, the smoker forfeits a prorated portion of the deposit for each failure to attend or provide data.

In addition to using monetary incentives, the concerned investigator will want to do everything *within reason* to assure that subjects do not drop out. There is a danger, however, of going overboard in an effort to control subject mortality. For example, if an experimenter were to call each subject before every treatment session, provide transportation to and from sessions, send thank you notes, and serve hot cocoa and cookies, subject mortality might be reduced, but the generality of the experimental results would also be limited to treatments that included the elements of phone calls, transportation, thank you notes, and refreshments. Since such procedural elements are confounded with the treatment, they must be considered an integral part of the treatment. Thus, when designing methods to control the drop-out problem, investigators should seek methods that can be used reasonably by therapists operating in other contexts.

Measurement Issues: Problems of Reliability and Validity

What is "smoking"? If the objective in smoking-cessation research is to reduce or eliminate "smoking behavior," then it is essential to define this target behavior precisely. If we cannot define it precisely, we cannot measure it reliably, and this means that we cannot possibly determine whether our interventions have had any effect on it.

Specifying the target behavior in measurable terms is not as easy as it may seem. What is the best unit of measure for smoking? For example, should we count the packs consumed, the cigarettes consumed, the puffs taken, the volume of smoke inhaled, or the amount of nicotine and tar ingested? Should we assess these monthly, weekly, daily, hourly, or by the minute? Should we take into consideration the various stimulus situations in which the behavior occurs?

The most common measurement unit is the number of cigarettes consumed per day—with no systematic classification of smoking situations. But there is nothing sacred about this particular unit; in fact, it ignores a number of potentially important variables, such as the number of puffs taken, the amount of smoke inhaled, or, for that matter, the fact that certain brands of cigarette are significantly longer than others. What if a subject lights a cigarette, smokes it halfway, extinguishes it, and relights it later—does that count as one or two? Despite its limitations, investigators should consider using this common unit of measurement in future studies—perhaps along with other units—simply because it provides a standard basis for cross-study comparisons. Whatever units are chosen, they must be specified in sufficient detail to permit other investigators to use precisely the same unit.

What method of measurement is best? Once the investigator has decided on a measurement unit, for example, counting the number of cigarettes consumed per day, then it is necessary to devise a suitable method of actually gathering the desired data. This practical requirement is another major source of difficulty in smoking research. The following is a partial list of measurement options avail-

able to the investigator, along with a brief discussion of possible advantages and disadvantages of each.

1. Laboratory methods. The most accurate method of measuring actual smoking behavior is to observe it under controlled conditions, such as in the laboratory. To achieve accuracy and control, however, the investigator inevitably sacrifices representativeness. That is, smoking behavior observed in the laboratory may bear little resemblance to unobserved, nonlaboratory smoking behavior. The appropriateness of using lab measures depends entirely on the specific experimental question. For example, if one were assessing the relationship between anxiety and smoking, it would be appropriate to begin doing so in the lab by systematically manipulating levels of stress while measuring the smoking behavior. However, if one were interested in the therapeutic value of a particular intervention, then changes in laboratory behavior would not be regarded as meaningful or persuasive evidence of therapeutic change. Smoking-cessation studies may include laboratory measures as part of a larger group of dependent measures, but there remains the problem of devising suitable extralaboratory criterion measures.

2. Self-report methods. This is the most commonly used assessment method in smoking studies. Subjects are enlisted as collaborators in the data-collection process; they are asked to monitor, record, and report on their own smoking behavior. Of course, the advantage of this method is that no one is in a better position to observe a person's smoking behavior—across all situations and at all times—than that person herself/himself. One disadvantage is that the person's self-reported data may be biased, inaccurate, or falsified, and thus there remains the need for a suitable independent measure of the subject's smoking behavior. Another possible disadvantage of self-report measures is that they sometimes can be reactive; that is, when subjects self-monitor their behavior, this may significantly affect the behavior being monitored in some manner (McFall, 1970). For example, it is common for subjects in smoking-cessation programs who are asked to self-monitor their smoking frequency during a

baseline period to report that the monitoring makes it difficult for them to continue smoking "normally." Nevertheless, because subjects do have unique access to their own behavior, it seems that the advantages of the self-report method usually outweigh its disadvantages—at least in smoking-cessation research—and that it will continue to be the principal data-collection method in such research. The problems with the method simply will have to be controlled or minimized as much as possible (e.g., see Nelson, 1977, for discussion of self-monitoring effects and their control).

3. Unobtrusive naturalistic measurement. Webb, Campbell, Schwartz, and Sechrest (1966) have suggested a variety of approaches that investigators might use in their effort to obtain useful naturalistic data without being so obtrusive as to contaminate the data. Translating their general suggestions into assessment methods for smoking behavior will require creativity and inventiveness, but one illustrative possibility is presented here to help stimulate the reader's own imagination:

Smoking ordinarily results in the accumulation of residual evidence in the form of cigarette butts. By monitoring a sample of the likely deposit sites of butts—for example, ashtrays in the office, home, or auto—an investigator might get a reasonably good picture of within-subject changes in smoking patterns over time. The resulting data should provide an indirect check on the accuracy of a subject's self-report.

There are at least three problems with unobtrusive naturalistic methods. First, there are ethical and legal problems with poking around in another person's personal space, such as their auto, home, or office, without their informed consent; but to obtain full consent would surely do away with the unobtrusiveness of the measurement. Second, the availability of naturalistic smoking data will vary from subject to subject, depending on the particular environmental settings that each subject regularly frequents; thus, the samples obtained for different subjects may not be sufficiently comparable to permit an analysis of group data. Third, the collection of unobtrusive naturalistic data may be prohibitively difficult or expensive.

4. Collaborator reports. If the subject's self-reports are suspect and if the investigator cannot arrange to observe firsthand the subject's naturalistic smoking behavior, then a compromise solution may be possible. Perhaps a third party—someone living or working closely with the subject—could be enlisted as an observer of the subject's smoking behavior. This assessment method has been used with increasing frequency in recent years. Investigators typically have asked subjects to provide names of persons who would be in a position to observe their smoking and who could be contacted periodically for reports.

One problem with relying on such collaborator reports is that the persons providing the data ordinarily are close friends of the subjects and thus are not necessarily any more objective reporters than the subjects themselves. It has not been uncommon, for example, for subject and collaborator reports to be extraordinarily highly correlated (e.g., over .95). Such agreement cannot be automatically accepted as evidence that the subjects and collaborators have provided valid data; the high correlation may reflect little more than collusion between the subjects and collaborators. Unfortunately, there is no simple method of assessing the validity of collaborator-reported data, which means that this measurement method cannot stand alone as a validity check on subjects' self-reports.²

5. Correlates of smoking behavior. For years, investigators have searched for a reliable correlate of smoking behavior that they could use as a sensitive indirect measure in their smoking studies. Nicotinic acid stains on the fingers, respiratory flow volume, sputum samples, and blood assays are among the various measures that have been considered at various times with only limited success. Recently, however, a promising physiological correlate of smoking behavior—carbon monoxide levels in samples of expired air—has been identified and used successfully in smoking studies (Danaher, Lichtenstein, & Sullivan, in press; Lando, 1975). Brockway (in press) has reported that thiocyanate may prove to be yet another objective measure of cigarette smoking. More work along these lines seems as though it might be helpful.

In summary, no single measure of smoking

behavior is adequate. Until the absolute or ultimate measure has been discovered, investigators must rely on a network of measures, each of which can serve to cover the weaknesses or blind spots of the others. In any event, a more convincing argument for validity can be made when there is concurrence among several independently derived measures.

Assessing change. Smoking-cessation studies typically are composed of four assessment periods: (a) a baseline period, during which subjects' pretreatment smoking behavior is recorded; (b) a treatment period, which can be broken down into several subperiods corresponding to different phases of treatment or to units of time; (c) the end of treatment; and (d) a follow-up period, ideally covering a minimum of 6 months to 1 year. By assessing changes in smoking behavior over these periods, it is possible to evaluate the effects of different interventions. However, to the extent that any of the assessment periods are not adequately designed and controlled, the meaningfulness of the results will be seriously limited. Some of the most common design problems are discussed below.

The essential requirement of the baseline period is that it provide a solid anchor against which to weigh the magnitude and significance of changes in any subsequent periods. Thus, the most serious mistake that investigators can make during the baseline period is to fail to assure that their measure of pretreatment smoking behavior has stabilized before they initiate the treatment period. This flaw is not always apparent from post hoc inspection of the published data. It has become common practice to report only one data point—usually the mean smoking frequency—for the entire baseline period; this practice is unfortunate. An absolute minimum of three data points is necessary if one wants to estimate the stability of baseline behavior. A single data point may obscure underlying trends in the data that might sub-

² It may be useful to distinguish between differences in accuracy as a function of the type of data being collected. It probably would be easier for a collaborator to report accurately on a subject's abstinence than on the subject's smoking rate.

stantially affect how one would interpret the experimental results.

Another common methodological problem encountered in the baseline period is that experimental groups sometimes are found to differ prior to treatment.¹ This difference need not be statistically significant to be considered serious. And when initial group differences do exist, there is little that can be done to correct the problem once the experiment has been carried out. Thus, before proceeding to introduce any differential treatments, investigators routinely should examine their baseline data to be certain that the groups are comparable. If the groups are not comparable, they still can be reconstituted at that point and the experiment can be salvaged. (After randomly reassigning subjects to treatment conditions, the baseline period must be repeated, of course.)

The treatment period obviously is devoted to the introduction of the experimental intervention. The introduction of the independent variable, however, does not mean that the dependent variable can be ignored during this period. On the contrary, it is essential to an analysis and interpretation of any experimental effects that the dependent variable be carefully assessed during treatment as well as before and after. Only in this way will it be possible to evaluate closely how the treatment exerted its effect on smoking. For example, in a smoking treatment designed to produce a gradual withdrawal from smoking, one would want to have several subjects actually follow the expected gradual decrease. In other words, assessments conducted during the treatment period permit an interim analysis of the treatment process itself, whereby adjustments can be made before and after treatment period on a basis of (1) individual and (2) group differences.

Lehmann (1971) has recently argued that there is too often an ad hoc analysis of the treatment process when the experimental researcher fails to produce a significant outcome difference. Subject to the foregoing, it is a sensible response that argument of design is better. That is, more effective treatments need to be designed for publication by the kind analysis of significant within-treatment group differences that were intro-

lated to meaningful outcome variables. Such analyses are potentially valuable, however, to the individual investigator who needs to understand what went wrong or what might account for the failure to obtain significant outcome differences. Without such process information, it is difficult to rise above one's failures and to design better treatments.

The bottom line question in smoking cessation research is the outcome question: Did the treatment work? This must be answered within two outcome time frames: immediate and long term. The end-of-treatment assessment provides an immediate measure of treatment effects; it also marks the transition from treatment to follow-up periods. When compared to the baseline measure, it provides a summary assessment of change over the treatment period. It also represents a benchmark against which to compare subsequent assessments and to evaluate the long-term maintenance of changes. The end-of-treatment assessment must contain the same measures used in the baseline period and in the follow-up period; otherwise, a valid assessment of change is not possible. That is, subtle variations in measurement procedures from period to period may make meaningful comparisons difficult or impossible. For instance, it would be questionable practice to assess pretreatment to posttreatment change by using self-monitored data at the baseline period and subjective estimates of smoking frequency at the posttreatment period.

A well-designed follow-up period is the one question of a valid therapy outcome study. Unfortunately, the follow-up period seems to be one of the weakest links in most smoking cessation studies. It seems such a waste of resources for an investigator to be meticulous about carrying out the baseline treatment and posttreatment periods, only to be lax about the follow-up.

Subject mortality is the most common follow-up problem in smoking cessation studies. It is a problem that seems to be repeatedly related to the length of the follow-up period. The longer the period (a virtue), the more likely that subjects will be lost (a fault). The unfortunate implications of subject mortality already have been discussed. The solutions are not as easily outlined. Perhaps the single

important factor in eliminating the problem of the persistence and determination of the investigator in pursuing the study of each every subject. When an investigator reports that some subjects "could not be located for the follow-up assessment, one cannot help but wonder to what lengths the investigator actually went to locate the missing subjects."

The follow-up measure of smoking should be comparable to the one used in the baseline and end-of-treatment assessments. It has not been uncommon for investigators to rely on unconfirmed global self-reports of smoking frequency, obtained from subjects via either telephone conversations or preaddressed postcards as their primary follow-up measure. Such a casual approach to assessment would be unacceptable in the other experimental periods. It is equally unacceptable in the experimental period from which ultimate conclusions concerning the treatment outcomes are drawn.

As has been pointed out elsewhere (McFall & Hammen, 1971), virtually any plausible smoking-cessation treatment that one can imagine is capable of producing a significant temporary reduction in smoking behavior. However, few treatments have managed to produce sustained reductions exceeding those obtained by placebo treatments or minimal-treatment control conditions. An assessment of changes in smoking behavior across the experimental periods, therefore, is bound to yield a statistically significant main effect in virtually all studies, but is unlikely to yield significant between-treatment differences. Only the discovery of significant treatment differences is remarkable enough, at this stage in the history of smoking cessation research, to warrant publication or general dissemination of the results. There are two exceptions to this rule. (One is when an established intervention that had come to be expected to yield significant results fails to do so, another is when a truly novel treatment approach, derived from a reasonably prominent theory, is found to have an effect. In general, the publication value or newsworthiness of a finding is directly related to its unexpectedness.

Interpretation of results. If the results of the smoking-cessation experiments are to be compared and integrated, they must be

presented in a manner that permits cross-study comparisons. In the past, many investigators have tended to report their results in unconventional ways, which has frustrated efforts to make such comparisons. A standardized format for data presentation would greatly improve this situation. (Of course, the use of a standard format would not prevent authors from presenting their data in other formats, in addition, when it served their special purposes.)

Convention suggests that the standard format might include as a minimum (a) period-by-period changes in smoking frequency, expressed in terms of a percentage of the baseline mean, and (b) the percentage of subjects within groups achieving total abstinence (and other lesser target levels, if appropriate). Both summary statistics should take into account all subjects who entered treatment, not just those who completed treatment.

Treatment Issues: Problems of Relevance and Replicability

Research strategies. Smoking cessation research can be characterized as "a problem in search of a cure." That is, investigators typically have been concerned with testing the effects of one or more experimental treatments on the smoking behavior of their subjects; the aim has been to find an effective method for helping people kick the smoking habit. The specific research strategy used in any given study usually falls into one of the following four categories.

1. **House races.** The most common strategy has been to line up several promising treatment methods along with a control treatment, and to give them all a run for the money under common conditions, with randomly assigned subjects from the same population. The judged by the same outcome measures. The competing treatments need not be related to one another in any systematic way other than their common goal. At the conclusion of the house race, the results are interpreted in a straightforward manner, namely in terms of the relative effectiveness of the various treat-

The new investigator might anticipate the over-take problem at the outset by obtaining names and addresses of persons who would know the whereabouts of each subject at follow-up.

ments included in that particular comparison. The results have little theoretical significance; they reveal little about the reasons why things turned out as they did.

2. *Dismantling strategy.* Once an effective change technique has been found, it can be examined more closely in subsequent studies in which it is systematically dismantled to see how its various components contributed to the overall treatment effect. This strategy has been used relatively infrequently in smoking-cessation research, to date, because few treatments have proved themselves sufficiently promising to warrant such an internal analysis. One exception has been the "rapid smoking with warm, smoky air" technique reported by Lichtenstein and his colleagues (Schmahl, Lichtenstein, & Harris, 1972). The experience of these investigators, however, illustrates one of the potential frustrations of the dismantling strategy. When Lichtenstein, Harris, Birchler, Wahl, and Schmahl (1973) compared the relative effectiveness of the rapid-smoking component, the warm, smoky air component, and the combined components, they found that all three treatment conditions were comparably effective. That is, each component alone yielded the same level of effect as the two components combined. In this instance, the dismantling strategy failed to reveal very much about the mechanics of the treatment effects.

3. *Constructive strategy.* This approach, like the preceding one, assumes that a fairly effective intervention has been found. Using the established intervention as a solid treatment base and as a comparison condition, additional components are systematically added in an attempt to increase the consistency or magnitude of the overall effect. The primary aim of this pragmatic strategy is to build a maximally effective treatment. To use this strategy, it is not necessary to understand precisely how the various components function or interact; this approach emphasizes outcome over process. However, without some guiding theory, the selection of components to be added to the treatment package is likely to be haphazard and inefficient.

4. *Theoretical research.* The most advanced research approach would be one based on a theory of smoking behavior and would be designed to test hypotheses derived from that

theory. Ideally, such research would lead to the most substantial advances, both scientific and technological. Unfortunately, there are few promising theories available at present to guide or stimulate such high-level research in the area of smoking behavior. History has shown that good theory is unlikely to come from the grand speculations of armchair thinkers; it usually arises out of attempts to integrate and organize empirical observations. Thus, there is reason to hope that smoking research eventually will manage to bootstrap its way from horse race studies, through dismantling and constructive studies, to theoretically grounded research.

An alternative research strategy deserves mentioning. It is a relatively unexplored, more indirect approach to the task of discovering an effective treatment. Investigators might learn a great deal if they took time out from their treatment studies to look closely at the various methods that have been successfully used by the multitude of former smokers who have quit on their own. There may be effective "folk methods" that could teach us a great deal about smoking behavior and its treatment.

Interpreting the results. An investigator's choice of a research strategy has important implications for the subsequent interpretation of the research results. There seems to be an inherent antagonism between the two research objectives of relevance and replicability. On the one hand, designs that foster relevance, such as those used in clinical trials or in horse race studies, tend to be so non-specific and uncontrolled that they cannot be easily replicated. On the other hand, designs that foster replicability tend to be so tightly controlled and specific that they have limited immediate relevance to the clinical treatment setting.

At the former extreme would be a clinical study in which clients are treated over an extended period by a combination of procedures that can be described in only the most general terms. The clients who quit smoking represent genuine successes, but it would be virtually impossible to specify and reproduce all of the factors that contributed to their quitting.

At the opposite extreme would be laboratory analogue experiments in which unmoti-

rated subjects are led to believe that their smoking behavior is being studied for reasons unrelated to smoking reduction, in which they are exposed to a brief experimental manipulation, and in which changes in within-lab smoking behavior are the dependent variables. The results of such an analogue study cannot reasonably be interpreted as bearing directly on the clinical treatment of smoking. Campbell and Stanley (1963) have presented with great eloquence and clarity a detailed list of factors that must be experimentally controlled before an investigator can hope to interpret the results of any experimental or quasi-experimental study. A summary of their presentation is beyond the scope of this article. However, if there are any readers who have never read the Campbell and Stanley monograph, it really should be considered required reading before attempting to design any smoking studies. In fact, readers who have read it but have not reviewed it recently are strongly urged to do so.

Generally, it is helpful to remind ourselves that the aim of experimental research is the elimination of plausible rival alternative hypotheses. Viewed from this perspective, experiments never prove that particular hypothetical conceptions are true; rather, at best, a hypothesis gains in stature to the extent that it manages to survive rigorous experiments that are designed to disconfirm it, whereas competing hypotheses are disconfirmed by such experiments. In smoking-cessation research, for example, four of the most common competing hypotheses are:

1. Perhaps no systematic change even occurs.
2. If change does occur, perhaps the experimental treatment does not produce a greater change than do competing treatments.
3. If the experimental treatment produces the greatest change, perhaps the change is not durable.
4. If the experimental treatment is associated with the greatest and most durable changes, perhaps such changes can be explained more simply by uncontrolled factors aside from those that are part of the experimental treatment per se.

Even after these basic "null" hypotheses and "nonspecific factors" hypotheses have

been eliminated, and a treatment has been established as unquestionably effective, there remains the task of sorting through all of the various possible explanations for the mechanics of how the treatment works. Again, this is a process of eliminating alternatives until only a few are left standing. (The ideal of having only *one* alternative remaining is seldom if ever achieved.)

The interpretation of experimental results must always be presented with reference to the particular rival hypotheses that either were discredited or failed to be disconfirmed. Since no single experiment is likely to pare the list of alternative rival hypotheses down to a single survivor, all prominent surviving competitors should be recognized in the discussion and interpretation of a particular study. Suggestions for future studies designed to test the survivors are always welcome.

Negative side effects. Recent events in the area of smoking-cessation research have helped to emphasize the need to be sensitive to possible unanticipated negative side effects of our experimental treatments. Lichtenstein's rapid smoking technique has been one of the few treatments to achieve reasonably convincing and consistent effects—for example, abstinence rates after 6 months of 57% or more (Lichtenstein et al., 1973). However, this treatment recently was found to pose a potentially dangerous health hazard itself, especially if applied to patients with impaired coronary circulation (Hauser, 1974; Miller, Schilling, Logan, & Johnson, 1977). Fortunately, the major proponents of the rapid-smoking technique have been sensitive to the dangers and have urged that medical precautions become a standard preliminary to administering the treatment (Lichtenstein & Glasgow, 1977). Proponents of other treatments should show a similar sensitivity to the possible risks or costs of their interventions.

Going public. Once an efficient, safe, and effective treatment for cigarette smoking has been established through well-controlled research, there remains the problem of translating such an experimental treatment into a valid procedure for widespread use with the general public. Campbell and Stanley's (1963) discussion of external validity issues explores this problem thoroughly; however, some investigators apparently need to be re-

mindful of some of the most critical issues. For example, a recently published self-help book purports to offer the consumer an experimentally validated method for kicking the smoking habit. Assuming that the particular method may have been shown to be effective when administered in a controlled setting to selected subjects by trained therapists is not a sufficient basis for "going public" with the method in the form of a popular self-help book. Before going public in this way, the authors should demonstrate empirically that the method is effective with subjects who buy the book and self-administer the treatment (see Glasgow & Rosen, 1978)!

Conclusion

The methodological vagaries and pitfalls of conducting smoking-cessation experiments have been outlined and discussed. Where possible, suggestions for improved research were also offered. Generally, the standards for good design in smoking research are not different from the standards for research in other clinical areas. The very nature of smoking behavior, however, poses certain difficulties in attempts to achieve those design standards. Specifically, measurement problems have been a chronic weakness in smoking studies.

Despite the extensive efforts of numerous investigators over many years, cigarette smoking continues to be a major health problem. An efficient, safe, effective treatment for smoking behavior remains an elusive goal, although progress toward this end seems to have been made in recent years. Hopefully, the present article will help hasten the day when the goal is realized and when it will be possible to help large numbers of smokers kick the habit.

References

- Bernstein, D. A. Modification of smoking behavior: An evaluative review. *Psychological Bulletin*, 1969, 77, 418-440.
- Borgatta, E. F., & Evans, R. R. (Eds.). *Smoking, health, and behavior*. Chicago: Aldine, 1968.
- Brockway, B. S. Chemical validation of self-reported smoking rates. *Behavior Therapy*, in press.
- Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally, 1963.
- Danaher, B. G., Lichtenstein, E., & Sullivan, J. M. Comparative effects of rapid and normal smoking on heart rate and carboxyhemoglobin. *Journal of Consulting and Clinical Psychology*, in press.
- Glasgow, R. E., & Rosen, G. M. Behavioral bibliography: A review of self-help behavior therapy manuals. *Psychological Bulletin*, 1978, 85, 1-23.
- Hauser, R. Rapid smoking as a technique of behavior modification: Caution in selection of subjects. *Journal of Consulting and Clinical Psychology*, 1974, 42, 625.
- Hunt, W. A. (Ed.). *Learning mechanisms in smoking*. Chicago: Aldine, 1970.
- Lando, H. A. An objective check upon self-reported smoking levels: A preliminary report. *Behavior Therapy*, 1975, 6, 547-549.
- Lichtenstein, E. Modification of smoking behavior: Good designs—Ineffective treatments. *Journal of Consulting and Clinical Psychology*, 1971, 36, 163-166.
- Lichtenstein, E., & Danaher, B. G. Modification of smoking behavior: A critical analysis. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 3). New York: Academic Press, 1976.
- Lichtenstein, E., & Glasgow, R. E. Rapid smoking: Side effects and safeguards. *Journal of Consulting and Clinical Psychology*, 1977, 45, 815-821.
- Lichtenstein, E., Harris, D. E., Birchler, G. R., Wahl, J. M., & Schmahl, D. P. Comparison of rapid smoking, warm, smoky air, and attention placebo in the modification of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1973, 40, 92-98.
- McFall, R. M. Effects of self-monitoring on normal smoking behavior. *Journal of Consulting and Clinical Psychology*, 1970, 35, 135-142.
- McFall, R. M., & Hammen, C. L. Motivation, structure, and self-monitoring: Role of nonspecific factors in smoking reduction. *Journal of Consulting and Clinical Psychology*, 1971, 37, 80-86.
- Miller, L. C., Schilling, A. F., Logan, D. L., & Johnson, R. L. Potential hazards of rapid smoking as a technique for the modification of smoking behavior. *New England Journal of Medicine*, 1977, 297, 590-592.
- Nelson, R. O. Methodological issues in assessment via self-monitoring. In J. D. Cone & R. P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Schmahl, D. P., Lichtenstein, E., & Harris, D. E. Successful treatment of habitual smokers with warm, smoky air and rapid smoking. *Journal of Consulting and Clinical Psychology*, 1972, 38, 105-111.
- U.S. Public Health Service. *Smoking and health: Report of the Advisory Committee to the Surgeon General of the Public Health Service* (U.S. Public Health Service Publication No. 1103). Washington, D.C.: U.S. Government Printing Office, 1964.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. *Unobtrusive measures*. Chicago: Rand McNally, 1966.

Common Methodological Problems in Research on the Addictions

Peter E. Nathan and David Lansky
Rutgers—The State University

Among the common methodological problems in research on the addictions reviewed in this article are (a) selective, incomplete, or biased reviews of the body of prior research from which a study has arisen; (b) reliance on inadequate or incomplete diagnostic criteria in choosing subjects for study; (c) choice of inappropriate comparison groups for treatment outcome research; (d) use of inadequate alcoholic analogues when alcoholic subjects are unavailable; (e) failure to adequately account for treatment dropouts in analysis of treatment outcome data; (f) unwarranted choice of single-subject over group designs in addictions research and vice versa; (g) failure to ensure that comparably trained, equivalently committed therapists provide both experimental and control treatments in treatment outcome studies; (h) failure to ensure that patients in both experimental and control treatments receive treatments as therapist- and time-intensive; (i) failure to follow patients for adequate lengths of time posttreatment; (j) failure to provide for adequate, multidimensional treatment outcome measures tapping a full range of patient behavior; (k) failure to exercise restraint, scientific modesty, and criticality in reporting results of one's own research; and (l) failure to recognize important differences between statistical and clinical significance.

This article identifies common problems in research on the addictions. It also offers suggestions for remediating these methodological problems. The addictions considered in the article include alcoholism and the drug dependencies. Because the literature on alcoholism is much more extensive than that on drug addiction, most of the common errors reviewed are drawn from the alcoholism literature. Despite this focus, there is enough commonality between methods of research on alcoholism and drug dependence that criticism in one research area almost always has relevance to the other.

The first problems to be considered are those arising from inadequate, incomplete, or

biased reviews of relevant literatures. Next, methodological shortcomings of subject selection, research procedure, and data analysis are detailed. Consideration of problems in the presentation of results and their interpretation and discussion concludes the article.

Literature Review and Statement of the Problem

Although it might be considered gratuitous to begin this article by pointing out the impact that selective, incomplete, or biased reviews of research have on the impact of research findings, it is important, nonetheless, to recognize that the alcoholism researcher may be particularly liable to this temptation. Otherwise called "enlightened historical selectivity," this tendency comes easily to the addictions researcher who is forced to choose one position, from several, on such controversial matters as etiology (Sociocultural, genetic, metabolic/physiological, and behavioral views are foremost.) and treatment.

Preparation of this manuscript was facilitated by National Institute for Alcohol Abuse and Alcoholism Research Grant AA 00259 to the first author.

Requests for reprints should be sent to Peter E. Nathan, Department of Clinical Psychology, Graduate School of Applied and Professional Psychology, Busch Campus, P.O. Box 819, Rutgers—The State University, Piscataway, New Jersey 08854.

(Dynamic, behavioral, medical, and peer-group Alcoholics Anonymous treatments are available.) This breadth of available perspective has, in turn, been largely responsible for the proliferation and perpetuation of invalid "common truths" and unsubstantiated "old wives' tales" in this field.

Selective or biased reviews of the alcoholism literature have supported a variety of such "truths" through the years, including that alcoholism is a medical disorder (and that as a result, abstinence is the only appropriate treatment goal for alcoholics), that all alcoholics experience loss of control over their drinking, that "dry" alcoholics are the best therapists for other alcoholics, that alcoholics drink to reduce prevailing high levels of tension and anxiety, that alcoholics are oral-dependent individuals who have not come to terms with their needs for nurturance, and so on.

One area of alcoholism and drug addiction research—having to do with the etiology of alcohol and drug dependence—has been especially subject to the stultifying influence of selective biases on review and interpretation of research results. Since significant federal funding for alcoholism and drug research became available about a decade ago, unitary models of dependence have been largely disproven. Specifically, there do *not* appear to be characteristic personality patterns that differentiate drug abusers from nonabusers, there is clearly *not* a single route to alcohol or drug dependence, and alcoholics and non-alcoholics do *not* appear to metabolize ethanol in discernibly different ways at comparable levels of ingestion. Instead, the mechanism of dependence, the etiologic process, the personality structure of alcoholic and drug-dependent individuals, and their motivation for and response to treatment all depend on far more than intrapsychic or physiologic factors alone. To this end, the literature now supports a more sophisticated view, that of a complex individual system interacting with personal history and environmental factors to yield an addiction. Among the most clearly articulated of these new, more constructive points of view is the behavioral viewpoint on the addictions, well-stated by psychologists Miller and Eisler (1975):

Within a social-learning framework alcohol and drug abuse are viewed as socially acquired, learned behavior patterns maintained by numerous antecedent cues and consequent reinforcers that may be of a psychological, sociological, or physiological nature. Such factors as reduction in anxiety, increased social recognition and peer approval, enhanced ability to exhibit more varied, spontaneous social behavior, or the avoidance of physiological withdrawal symptoms may maintain substance abuse. (p. 5)

From another perspective, that of tongue-in-cheek, Mark Keller (1972), alcoholism's long-time Dean of Letters, says the following about unidimensional theories of alcoholism:

A splurge of reports, in the 1940's, of biochemical characteristics purporting to differentiate alcoholics from nonalcoholics stimulated me to review a voluminous related literature, implicating physical, social and psychological demarcators as well. The only conclusion I could derive, from the entirety of the reportage, took a form that became known, among colleagues, as Keller's Law: *The investigation of any trait in alcoholics will show that they have either more or less of it.* Accordingly, I then predicted that if sexadactly should be investigated, alcoholics will yield either more or fewer six-toed and six-fingered people than a control population. (p. 1147)

Subjects

Diagnostic criteria. The familiar problems that plague diagnostic decision makers called on to make diagnostic distinctions among psychiatric patients (Chapman & Chapman, 1977; Goldberg, 1968; Nathan, 1967) also confront the researcher who must select alcoholics from a general population and, as important, assure that the subjects chosen are representative of an identifiable portion of the universe of alcoholics. It might appear that choosing alcoholics must be easier than selecting schizophrenics or neurotics, because all alcoholics share a single stigma—they drink too much. But the task is made more difficult than it might seem by lack of consistency, in research reports, in detailing this hallmark of the disorder. For example, descriptions of alcoholics recently studied by psychologists range from "twenty subjects [who were] psychiatric patients with a primary diagnosis of alcoholism" (Levine & Zigler, 1976, p. 141), and "alcoholic Ss [subjects] [who] were volunteers from the patient population of the Alcoholism Treatment Program at the VA hospital" (O'Leary, Radford,

(Chaney, & Schau, 1977, p. 580) to

subjects [who] were three male veteran inpatient volunteers, ages 40, 50, and 55 years. They reported histories of chronic drinking of 15, 20, and 26 years, respectively. Each had experienced blackouts and delirium tremens as a result of drinking and had been hospitalized 9-20 times for alcohol-related problems. The subjects were each classified as binge drinkers consuming up to 140 ounces of beer and 40-70 ounces of liquor per day when drinking. Drinking episodes ranged from 5-30 days followed by a 10-day (on the average) period of sobriety. (Doleys, Caminero, Wallach, & Davidson, 1977, p. 207)

Leaving aside the question of which of these studies most carefully ensured that its subjects were in fact alcoholics, it is clear that only the latter article provides enough information to permit assessment of the extent to which the subjects studied were comparable to alcoholics studied elsewhere.

The greatest possible detail about drinking pattern must be conveyed in a research report. Above all, researchers—and their audience—must become unwilling to consider men and women as alcoholics simply because they received that diagnosis from someone somewhere at some time. Instead, hard signs of physical dependence, psychological dependence, and tolerance ought to be in evidence, along with indications that alcohol has been a problem in living and that problems in family relationships, vocational adjustment, and interpersonal relationships have followed from excessive alcohol consumption.

Although use of a broad range of assessment procedures is an essential aspect of alcohol and drug treatment research, the ultimate utility of these measures can only be gauged by reference to their respective reliability and validity estimates. It is, therefore, incumbent on the responsible researcher to assess and report the adequacy of these measures as applied to the subject population in question. Unfortunately, it is rare that the reliability and validity of assessment procedures are reported in alcohol and drug treatment studies.¹

It is also important that age, sex, and socioeconomic status, at least, be summarized for all subject populations. As we have already observed, alcoholism does not wipe out the effect on behavior of subject variables

such as demography, education, and social class.

Even more formidable assessment problems confront the drug researcher, since only very recently have attempts been made to measure quantity and frequency of drug ingestion (e.g., Mirin, Meyer, McNamee, & McDougle, 1976; Rawlins, Randall, Meyer, McNamee, & Mirin, 1976). As a consequence, diagnosis of drug dependence is from whatever signs of physical dependence are available, whereas differentiation among varieties of drug abusers must be confined to descriptive typologies (e.g., heroin addict, amphetamine addict, polydrug abuser, and so on). However, Carlin and Stauss (1977) recently explored the possibility of categorizing drug addicts along functional dimensions, including a streetwise/straight typology (e.g., legality of primary means of support, conventionality of dress and grooming, ability to buy drugs from street drug dealers, etc.) and a self-medication/recreational drug-use typology. The major point of their article is that use of these descriptors, in conjunction with the more common descriptive typology, provides more information about subjects than the descriptive typology alone.

Control and comparison groups. Another common subject selection problem is choice of comparison groups when one wishes to contrast alcohol or drug abusers with nonabusers. If subjects are inpatients, must the comparison group also be hospitalized? And if it is, should it be composed of hospitalized psychiatric patients, because, presumably, alcoholics and drug abusers are often victims of psychiatric disorder as well? By the same token, if subjects are outpatients, can comparison subjects be chosen from any non-drug-abusing group or must they be psychiatric outpatients? In the same vein, how necessary is it to match experimental and control groups by age, sex, socioeconomic status (SES), level of education, and treatment motivation (if the latter is relevant to the study)? Our view

¹ For detailed information on unobtrusive approaches to the assessment of alcoholism and its concomitants, the reader can consult Briddell and Nathan (1976), Miller (1976) and Nathan and Briddell (1977).

of this complex matter is that alcoholics or drug addicts whose addictions are accompanied by concurrent psychiatric disorder must be compared to age-, sex-, and SES-matched nonaddicted individuals whose psychiatric diagnoses approximate those of the alcoholics. We do not assume, of course, that all drug-dependent individuals—or even most of them—are psychiatrically disordered just because they abuse drugs or alcohol.

Though an obvious and well-accepted control procedure, it is probably worth noting here, in passing, that groups of patients to be compared on the basis of treatment outcome must have been chosen from the same patient population to begin with, then assigned to treatment groups in a way guaranteed to ensure comparability. Of equal importance, in this context, is selection of a proper comparison treatment for alcoholics or drug abusers undergoing a nonstandard treatment whose efficacy is being assessed. In this regard, it is probably worth noting that "no treatment" or waiting list controls may be unethical when one is dealing with severely disabled addict populations. As important, they may also be unfeasible, since alternative treatment sources are often freely available to the motivated alcoholic or drug abuser. Most important, one must ensure that the control treatment that matched patients do receive is as active, long lasting, and time intensive as the new treatment being evaluated. This important issue is discussed in greater detail later.

A related control issue, in this context, is the treatment motivation of experimental and control subjects of comparative treatment studies. Failure to assess subjects' treatment motivation in treatment outcome studies involving small numbers of subjects, a common failing, prevents one from knowing whether apparent differences in treatment efficacy, if found, reflect real differences in the power of one treatment to effect changes in behavior or whether that treatment group simply included individuals who were more highly treatment motivated. Researchers fail to undertake this necessary assessment because doing so inevitably reduces the size of the pool of potential subjects and because the assessment is really not so simple. Above all, subjects may not tell the truth about their

treatment motivation, especially when financial, vocational, or judicial contingencies have brought them to treatment. But despite this problem, it is absolutely necessary to attempt to assess treatment motivation and to report results of that assessment in any study in which motivation for treatment could play a role in outcome.

Alcoholic analogues. Despite their omnipresence, alcoholics are not easy to locate for study. They are also not likely to be highly motivated to participate in psychological research (whose payoff to them may be unclear), are not always reliable in keeping appointments when they do agree to be subjects, and, when they are not drinking, they are apt to prefer working to participating in research. Because research designed to examine alcohol's effects on alcoholics can only include alcoholics in good physical and psychological health despite their chronic alcoholism, those who design such research find themselves with an even smaller pool of alcoholics ready, willing, and able to be research subjects.

For these reasons, researchers may choose to study *alcoholic analogue subjects*, who may be "problem drinkers" of one sort or another or, commonly, heavy-drinking college students. Though choice of such subjects is understandable, generalization from their behavior to that of alcoholics must be done with great care—if at all. For this reason, it is almost always preferable to study alcoholics than their analogues; it is practically impossible to control for all of the differences that separate alcoholics and nonalcoholics, some of which remain unknown. Among the most obvious of these differences are age, educational level, cognitive functioning, and rate of ethanol metabolism.

Above all, one must be chary about concluding that what is characteristic of the behavior of a group of problem or heavy drinkers is likely also to characterize the behavior of alcoholics—only more so! In all likelihood, such a conclusion is not justified.

The problem of treatment dropouts. Although the dropout problem for all kinds of psychological treatment is a serious one, it is especially so for alcoholics and drug addicts, both of whom are poorly motivated for treatment; variable in meeting vocational, familial,

and personal obligations; and, often, in poor control of their behavior. A recent study of more than 14,000 chronic alcoholics treated at 44 federally funded alcoholism treatment centers throughout the country (Armor, Polich, & Stambul, 1976) revealed that fewer than 15% of these clients continued in treatment as long as 3 consecutive months. Moreover, of the original sample of 14,000, 6-month follow-up data on only 2,371 clients were reported. Both figures suggest something of the scope of the attrition problem facing the addictions treatment outcome researcher.

Given, then, that substantial numbers of alcoholics and drug addicts drop from treatment in its midst and that as many more who complete treatment cannot be located for follow-up, how is one to compare two or more treatment methods on both the short and long term? To begin with, one must include in any statistical analysis of differences among treatment groups subjects who dropped from treatment; these subjects should be considered treatment failures regardless of the rationalizations some may have given for the decision to terminate. The unintentional deception that can result if this rule is not followed is illustrated by an article published a few years ago comparing groups of alcoholics receiving two "active" treatments, both involving the administration of electric shock, and two "inactive" placebo treatments. On comparing the four groups in terms of average length of abstinence immediately following treatment, the study's authors concluded that one of the active treatments was associated with significantly longer periods of abstinence than the other three treatments. Inspection of the data on which this conclusion was drawn, however, reveals that 13 active treatment subjects dropped from treatment in its midst, whereas only 1 placebo treatment subject did so; this fact was not taken into account in the significance testing of posttreatment abstinence rates, which only compared subjects who had completed treatment. Since many of the active treatment subjects dropped from treatment because they found electric shock aversive, to conclude from the abstinence rates of subjects who were motivated enough to endure shock and remain in treatment that the active treatment

was more effective than the placebo is deceptive, however unintentional.

Although most readers doubtless agree that patients who drop from treatment must be considered treatment failures for comparative purposes, it is more difficult to find agreement on how to handle treated subjects who cannot be located for follow-up. Is one to conclude that all subjects who cannot be found at a follow-up interval have returned to drinking or drug taking, have remained "dry" or "clean," or are using alcohol and drugs in about the same proportions as subjects who could be located? Though all three stances are defensible, the most conservative approach to this problem is to consider all patients unlocated at follow-up to be treatment failures (a position recommended by the American Medical Association, 1956; Pattison, Sobell, & Sobell, 1977; Sobell, 1978).

Before having to accept this admittedly unsatisfactory solution to the problem, though, one ought to take every step necessary to ensure that as many patients as possible are available for follow-up assessment. Ways to achieve this desirable goal are summarized later in this article in the section entitled *Follow-up*.

Who to study? It is inevitable that most of the subjects in alcoholism and drug treatment research will be male skid-row or "blue collar" alcoholics and ghetto-dwelling drug addicts. The reasons are that these individuals are often unemployed and, hence, available for intensive, long-term study; even if employed, they are more likely to be clients of publically funded clinics whose clientele is more available to researchers than are the patients of private psychologists and psychiatrists. Another is that this population of alcoholics and drug abusers is more likely than any other to reach the attention of the judicial system, and subsequently to be remanded to treatment by the courts; such court referrals constitute an important source of subjects for treatment evaluation studies. This is especially important, since it is clear that data on treatment outcome for court-referred drug addicts—whatever their socioeconomic status—cannot readily be generalized to drug abusers referred for different reasons. Unfortunately, few alcohol or drug

treatment reports include explicit descriptions of referral sources.

As a result of these and other similar factors, most of our data on drinking patterns, patterns of drug use, and response to treatment are from this group of subjects, even though it is not representative of the total universe of drug abusers and alcoholics. It has been estimated, for example, that fewer than 5% of all alcoholics are of the skid-row variety (Armor et al., 1976).

With these observations in mind, we offer two remedial suggestions: (a) Every effort should be made to study socially and economically advantaged alcoholics and drug abusers as well as more readily available groups. Sources of such persons include autonomous Alcoholics Anonymous groups; private clinicians who might agree to cooperate with researchers they know and trust; and private hospital administrators, whose interest in research collaboration might reflect their wish to gain professional credibility by associating with a university's research efforts. (b) If the usual groups of skid-row alcoholics and destitute heroin addicts are the only populations available for study, the researcher must make clear that generalizations from his/her data can only be to comparable groups of intellectually and vocationally limited individuals, that few alcoholics and drug-dependent individuals are as impoverished in so many ways as are these overstudied individuals.

Procedure: Comparative Treatment Studies

Group versus single-subject designs. The same design dilemma faces the alcoholism or drug treatment researcher as confronts any other clinical researcher who must choose between group and single-subject designs. Group designs permit more comfortable generalization of findings from one population to another, because they enable statistical testing of differences in the efficacy of one treatment procedure and another among substantial groups of individuals. At the same time, group designs sacrifice the fine-grained analysis of the temporal dynamics of behavior change provided by single-subject designs. As important, they do not permit the "return to

baseline" implicit in the single-subject ABA design that is necessary to establish that it was the treatment, rather than time, motivation, or expectancy, that brought about the observed changes in behavior.

Because so much of the research on alcoholism treatment by psychologists during the past 5 years has been behavioral, much of the literature reports single-subject designs. By contrast, the older literature reported results predominantly from group designs. It is possible, in fact, to retain the virtues of both designs by ensuring that group designs provide for thorough and reliable pretreatment and posttreatment assessment of behavior, carefully matched experimental and control treatment subjects, and suitably prolonged follow-ups.

Several recent studies comparing alcoholism treatment methods have achieved this desirable blending of the strengths of single-subject and group designs, among them comparisons of abstinence-oriented and controlled-drinking-oriented behavior modification programs (Sobell & Sobell, 1976), behavioral family counseling, electrical aversion, covert sensitization, and systematic desensitization (Hedberg & Campbell, 1974), and two broad-spectrum behavioral treatment packages (Vogler, Compton, & Weissbach, 1975). All three projects provided for comprehensive pretreatment and posttreatment assessment of drinking behavior and vocational, familial, and interpersonal effectiveness; follow-up periods extending to or beyond a year, accompanied by procedures to minimize subject attrition during the follow-up period; treatment groups matched for relevant demographic and treatment motivation variables; and careful attention to comparability of "active" and "placebo" treatments. Unfortunately, the blending of constructive design elements contained in these studies is not generally characteristic of the field.

Who are the therapists? It is important to ensure that the treatments contrasted in comparative treatment studies be provided by comparably trained, equally committed therapists. To compare treatment outcomes when one treatment is given by experienced therapists and the other by graduate students, or when one is administered by clinicians com-

mitted to the treatment they are providing while the other is offered by men and women unconvinced that what they are doing has value, represents poor research design. Yet many researchers have made precisely this design error when comparing one or another combination of innovative therapeutic approaches with what is euphemistically termed *standard hospital milieu therapy*. The latter, which may include alcohol or drug education, group therapy, occasional personal counseling, and the opportunity to participate in an Alcoholics Anonymous or Synanon group, is usually provided by undertrained, underpaid state hospital workers whose enthusiasm for their work, insight into its value, and level of clinical training are rarely equal to those administering the new treatment approach. When the new therapeutic package turns out to be more efficacious than the standard one in such comparative studies, one cannot be certain that it was actually more active or that it was provided by committed therapists whose enthusiasm for their work was infectious.

We recommend that no new therapeutic package be compared to "standard hospital treatment milieu," so often a euphemism for virtually no treatment at all. Instead, the separate components of novel treatment packages might more profitably be compared to each other, as well as to the package as a whole, in this way permitting assessment of each component's contribution to the overall package's effectiveness. For those researchers wishing to compare their efforts to some "standard" success rate for alcoholism treatment, two choices are among those possible: Armor et al.'s (1976) recent survey data, suggesting that 70% of the clients treated at typical alcoholism clinics across the country show short-term improvement in drinking rate, make for rigorous comparisons; Ditman's (1967) earlier conclusion that the success of Alcoholics Anonymous, often the alcoholic's initial treatment resource, is only between 30% and 35% makes for much more comfortable ones. Our own view of the matter is that abstinence beyond a year by 40% or more of patients treated by *any* technique surpasses current expectations.

Treatment variables. To justify meaning-

ful comparisons among different treatments, it is axiomatic (a) that all patients in all treatment groups must have received about the same number of hours of individual and group treatment and (b) that the intervals between treatment sessions must have been comparable. It is necessary to ensure comparability of treatments in this way because there is a considerable clinical treatment literature that attests to the therapeutic impact of the relationship between therapist and patient, an impact independent of the therapeutic methods that the therapist chooses to use (Smith & Glass, 1977).

In practice, this straightforward control procedure can present problems. Comparing an experimental treatment to "standard hospital milieu" treatment, for example, might prove impossible if this control is taken seriously; Most experimental treatment programs provide individual or group treatment—or both—administered by highly trained and experienced clinical researchers and their students, whereas standard treatment, customarily administered by hospital workers whose training and motivation may be inferior, is also likely to involve much less 1:1 contact and small group treatment. In similar fashion, comparison of methadone maintenance to psychotherapy is frequently confounded, since methadone programs usually involve the drug addict in less treatment time and less intimate therapist contact than do most psychotherapeutic treatments.

A related matter has to do with the scope of treatment issues addressed—the range of problem behaviors confronted. Consider how different the scope of treatment was for experimental and control subjects in a recently reported, widely cited behavioral treatment package for alcoholics:

Experimental Subjects: Procedures included subjects being videotaped while intoxicated under experimental conditions, providing subjects when sober with videotape self-confrontation of their own drunken behaviors, shaping of appropriate controlled drinking or nondrinking behaviors . . . the availability of alcoholic beverages throughout treatment, and behavior change training sessions. [The latter] is a summary phrase to describe sessions which concentrated upon determining setting events for each subject's drinking, training the subject to generate a series of possible alternative responses to those

situations, to evaluate each of the delineated alternatives for potential short- and long-term consequences, and then to exercise the response which could be expected to incur the fewest self-destructive long-term consequences. Behavior change training sessions consisted of discussion, role playing, assertiveness training, role reversal or other appropriate behavioral techniques.

Control Subjects: Control subjects received conventional treatment procedures which could include group therapy, chemotherapy, Alcoholics Anonymous, physiotherapy and other traditional services. (Sobell & Sobell, 1973, p. 601)

Although the range and diversity of potential treatments available to control subjects in this study could have been as great as those provided to experimental subjects, one is left with the distinct impression that the scope of treatment offered to experimental subjects—the specific problem areas that the treatments were designed to confront—was both better targeted and more comprehensive than that available to control subjects. In other words, *the treatment package offered to experimental subjects was not only different from but better than that provided control subjects.*

One resolves this common design problem by confronting the fundamental difficulty of premature comparison of the efficacy of comprehensive treatment packages, by resigning oneself to the necessity to compare the components of a treatment package separately before attempting to assess the overall efficacy of the package itself. It was this strategy that guided Wilson, Leaf, and Nathan's (1975) assessment of the efficacy of electrical aversion, a common element of broad-spectrum behavioral treatment for alcoholism. Their subjects received lengthy conditioning trials of electrical aversion, followed directly by ad libitum access to beverage alcohol. When subjects showed absolutely no evidence of conditioned aversion to ethanol, by drinking during the posttreatment period with the same enthusiasm for alcohol that they had shown pretreatment, it was concluded that electrical aversion, either by itself or within a broad-spectrum treatment package, did not have the therapeutic power many had presumed it to have.

Follow-up. Although it is obvious that the longer the follow-up interval posttreatment,

the more sure one can be of data on treatment efficacy, no one knows how *short* the follow-up period can be and still reflect ultimate treatment efficacy. Nonetheless, most alcohol and drug researchers question any follow-up interval that fails to extend to 1 year or more posttreatment. They believe that such an interval is inadequate to assess the long-term effects of treatment for alcoholism or drug dependence, given that many alcoholics and drug abusers spontaneously modify or cease drug or alcohol ingestion on their own for periods extending beyond a year. We believe that a 2-year follow-up of treatment for alcoholism or drug dependence is necessary to provide a suitably comprehensive view of the power of the treatment to effect lasting change.

A 2-year follow-up period, however, presents formidable subject attrition problems. Given the transient nature of the existence of many alcoholics and drug addicts, how does one keep contact over that time with men and women whose residences, jobs, and lives change so much and so often? Sobell (1978), a clinical researcher with great experience in maintaining contact with alcoholic clients over lengthy follow-up intervals, suggests the following set of coordinated steps to minimize attrition of subjects during a follow-up period: (a) Allow enough time and develop enough persistence to locate as many follow-up subjects as possible; do not settle for a majority of subjects or even for most of them. (b) Explain to subjects at the end of treatment why follow-up contacts are scheduled, when they can expect to be contacted for follow-up, the kinds of information to be requested, and how this information will be used. (c) Identify as many collateral sources of information about subjects as possible during treatment. (d) Maintain continuity of contact with patients during the follow-up interval by keeping in touch every few weeks, even if follow-up information is not required until 6-month or 1-year marks. (e) Be prepared to consult official records (e.g., jail, hospital, welfare, driver records, etc.) to locate lost subjects. Parenthetically, maintaining this kind of frequent contact for follow-up purposes also serves as an important source of low-cost "continuing care"

that may help maintain therapeutic gains first achieved during treatment. In this instance, then, good research design and good patient care go hand in hand.

A frequently ignored follow-up issue of relevance to both alcohol and drug treatment studies is that a treatment program may be highly effective in attaining desired goals while patients are actively involved in the program, only to appear to fail when patients return to nonsupportive or destructive environments. Unfortunately, a follow-up assessment 3 or 6 months posttreatment will not reveal this Treatment \times Environment interaction (Gotestam, Melin, & Öst, 1976). Under these circumstances, assessment of outcome immediately after treatment has ended and then again during a follow-up period much more clearly delineates dynamics of improvement and determinants of relapse.

Outcome measures. A variety of direct and self-report measures of drinking behavior have been developed. The Cahalan quantity-frequency index (Cahalan & Cisin, 1968), the Michigan Alcoholism Screening Test (MAST; see Selzer, Vinokur, & Wilson, 1977), and Marlatt's behaviorally oriented Drinking History questionnaire (Marlatt, 1975) are the most widely used self-report measures of drinking. Although use of such measures opens the investigator to criticism on grounds of the relative unreliability of self-report data, several studies have recently reported that alcoholics' and drug addicts' self-reports on drinking and drug ingestion may be far more accurate than previously believed (Cox & Longwell, 1974; Sobell, 1978; Homer & Ross, Note 1). Regardless of these data, however, we prefer direct measures of drinking behavior, including blood alcohol determinations taken at unannounced intervals (e.g., Miller, 1975), taste test or other drinking analogue methods (e.g., Marlatt, Demming, & Reid, 1973), or, simply, ad libitum drinking opportunities (Schaefer, Sobell, & Mills, 1971). The latter two methods, however, cannot be used to assess dry alcoholics, since they provide the opportunity for alcohol ingestion; this assessment problem is considered in greater detail by Nathan and Lansky (1978). Whatever outcome measures are ultimately selected for use, the point discussed earlier in regard to

the issues of reliability and validity is relevant here as well: These measures are relatively uninterpretable unless reliability and validity estimates for the population under study are reported.

There are other measures of treatment outcome that relate less directly to drinking or drug ingestion. One of the most thorough tests of these pretreatment and posttreatment measures was provided by Armor et al.'s (1976) recent national study of 44 alcoholism treatment centers scattered around the country. A complex of vocational, marital, and social indicators of relevance to alcoholism were tapped; direct reports from employers, clinicians, relatives, and others in a position to comment on job stability and marital adjustment were elicited. In some instances, these indices of behavior change showed more dramatic improvement following treatment than did direct measures of change in alcohol consumption.

In similar fashion, treatment goals for drug abusers extend far beyond mere reduction of the frequency of drug abuse; goals are often set to include improved employment status, widened spheres of non-drug-related social contacts, and improved physical health (Anderson & Nutter, 1975; McCabe, Kurland, & Sullivan, 1975). As a result, pretreatment and posttreatment measures must tap information that directly reflects the status of these socially desirous variables, frequently a most difficult task.

Whichever combination of direct and indirect indicators of change-in-life functioning is chosen, it is important that follow-up intervals be circumscribed enough to permit reliable recall of the data requested. To ask an alcoholic or drug addict to recollect quantity and frequency of alcohol or drug ingestion or quality of marital or vocational adjustment over a 6-month or 1-year period with any reliability is obviously impossible; to request such recall for a 2-week period is both more reasonable and, likely, will prove more reliable.

Procedure: Other Studies

Nature of the analogue in analogue studies. One of the most common thrusts of alcohol-

ism research that is not treatment oriented involves the use of alcoholic analogues. Such research might describe "typical" drinking patterns of chronic alcoholics, college students, or middle-class women. It might investigate the effects of alcohol on psychological, social, or cognitive functioning. Or the research could inquire into the impact of one or another environmental variables, designed to induce stress or anger, for example, on consequent drinking. All of this research could be undertaken analogically. Typical drinking in the real world can be inferred from patterns of drinking in the laboratory; under certain circumstances, the effects of large amounts of alcohol on cognitive functioning can be predicted from the effects of small doses; in some cases, the impact of stress on drinking can be studied by first creating artificial stress in the laboratory—an environmental condition analogous to real-life drinking stress—and then measuring consequent drinking behavior.

Reasons for undertaking analogue research instead of research in the real world are many. Alcoholics and drug addicts are hard to find, even harder to study. Giving alcohol in large doses to anyone, especially to an alcoholic, is difficult; giving hard drugs to anyone, especially to a drug addict, is virtually impossible. For these and other reasons, analogue research on the addictions stands between no research and the researcher. But there are important caveats to observe in undertaking analogue research in this field. They include, above all, assuring oneself—and one's professional audience—that the analogue one has chosen to use is not stretched so far as to bear only vague resemblance to the phenomenon in the real world. For instance, the laboratory drinking behavior of college sophomores cannot be considered representative of the drinking of adults in neighborhood taverns or at home. Similarly, the absence of an impact on intelligence test performance of moderate blood alcohol levels during a single test session has little relevance to possible cognitive deficits suffered by long-term heavy drinkers. Finally, stress induced in a laboratory might not be at all comparable to stress in the real world, especially when the latter

stress derives from *things that matter* and the former stress does not.

The obvious solution to these potential problems is to avoid analogue studies, when possible. But when it is not possible to do so, as is so often the case, what then? In that event, one must document, to the extent possible, the nature of the relationship between the analogue and the real world. The burden of proof of the relevance of an analogue to the natural environment is on the investigator, not his or her audience!

Experimenter bias. By now virtually every psychologist knows of the pitfalls of experimenter bias, known to many as the "Rosenthal effect" (Rosenthal, 1966). In essence, experimenter bias refers to the unintentional bias that experimenters bring to their interactions with experimental subjects that can, in some cases, affect data.

Although it is doubtful that experimenter bias plays a more serious role in research on alcoholism and drug dependence than it does elsewhere, it can—and does—play a role that must be anticipated. Examples include the following: (a) In studies of free ad libitum drinking, especially those taking place in controlled laboratory settings, it is entirely possible for staff unintentionally to bias subjects toward either more or less drinking. This possibility is so real that we actively seek to reduce interaction between drinking subjects and staff to an absolute minimum in our own laboratory drinking studies. (b) In studies in which the effects of stressors on drinking or of alcohol on response to stressors is the focus, the investigator must ensure that the effects of the stressor derive from its real impact, not from his/her unintentional conveyance of expectations about those effects. In such studies, postexperimental questionnaires must affirm that the subject did perceive the stress stimulus as a stressor. (c) In studies in which alcohol's effect on other behaviors is being examined (e.g., its impact on projective responses, interpersonal facility, or psychomotor behavior), it is just as important to ensure that hypothesized effects are not translated into observed effects, because the experimenters betray their expectations to drinking subjects. To this end, careful inquiry at the conclusion of the study as to

subjects' perception of its intent is frequently most enlightening.

The demography of alcoholism and drug dependence. As emphasized above, it is crucially important to control for variables other than variety of alcoholism or drug dependence in comparative treatment studies. Similar controls—or, at the least, awareness of the importance of such variables as age, sex, race, and socioeconomic status when experimental controls are impossible—are also necessary in studies exploring the impact of alcoholism on behavior. To fail to account for these variables is to run the risk of attributing to alcoholism or drug abuse responsibility for a particular behavior, a Rorschach percept, a Minnesota Multiphasic Personality Inventory (MMPI) or Wechsler Adult Intelligence Scale response pattern—or a treatment outcome—when, in fact, these behaviors may derive instead from the complex of ethnic, educational, or socioeconomic factors associated with alcoholism or drug dependence.

Results

On reporting results. It is tempting, when reporting the results of a study, to be optimistic about their significance, even a bit grandiose about their staying power. Human nature being what it is, it is hard to be overly critical of the natural human tendency to see in one's own efforts what one might not see in those of others.

On the other hand, although overestimation of the significance of a new nodule on the backside of a personality theory can be tolerated, misleading, immodest, or unrealistic reporting of the results of comparative treatment studies goes beyond bad form. When read by persons unequipped to separate an inflated claim from a barely significant probability value, such reporting could result in the misapplication of unproven methods.

The point we wish to make here is a simple one: It is incumbent on the serious researcher to report data parsimoniously, modestly, and completely, even when data in their entirety are more confusing and less unequivocally interpretable than data that have been selectively pruned. Although improper presentation of data is a snare for every re-

searcher, it is particularly so in fields such as alcoholism and drug dependence. In these fields, little is known; the stakes are high for persons who develop viable theories of etiology or useful therapeutic approaches; and rapid, independent confirmation or disconfirmation of new findings is difficult. As a consequence, a series of revolutionary findings have swept the field, including several genetic theories of etiology, a variety of new, "sure fire" diagnostic methods, novel behavioral treatment procedures that do not require abstinence, and innovative prevention programs guaranteed to reduce prevalence of these disorders. Unfortunately, like similar claims made of new discoveries about the schizophrenias in the late 1950s and early 1960s, the light of day—and independent replication—has either diminished or disproved most of these epochal findings.

Although Job was not necessarily speaking of psychological science when he concluded that there is nothing new under the sun, it might be wisest to heed his words until, or unless, independent confirmation supports your Nobel-prize-winning discovery, especially if you work in alcoholism or drug dependence!

Clinical versus statistical significance. It is as possible to overinterpret statistically significant as marginally or wholly insignificant differences among groups, because statistically significant differences may fail to achieve clinical significance. Although not conclusive, the following list includes three instances: (a) Psychological test patterns that differentiate groups at .05 or .01 levels rarely prove helpful in differentiating for diagnostic purposes. For example, alcoholics and drug abusers often differ from nonpsychiatric controls or unselected psychiatric patients by scoring higher on the Depression, Psychopathic Deviate, and Mania subscales of the MMPI; to draw similar diagnostic distinctions on the basis of symptoms of depression, mania, and psychopathic behavior alone, however, is rarely sufficient. (b) Statistically significant differences in the efficacy of one treatment approach over another, no matter what the level of difference in efficacy or how it was judged, are inadequate bases for selection of treatment unless or until cross-

validation of the promising approach, with different therapists and different patient groups and in different settings, confirms the general utility of the technique. Electrical aversion, a promising behavioral technique that made excellent theoretical sense and showed exciting promise when first applied, has since proven itself to be of little value as a therapeutic technique on extensive cross-validation. Methadone maintenance may be in the midst of suffering the same ignominious fate. (c) Studies demonstrating the significant impact of a specific etiologic factor on a specific disorder do not, perforce, prove that every patient possessed of that factor will develop the resultant disorder or that every patient carrying that diagnostic label can point to that etiologic factor. In the case of alcoholism, for example, recent research by Goodwin and Guze (1974) strongly suggests the role of genetic factors in the etiology of alcoholism. Significantly more Danish children of alcoholic parents given up for adoption developed alcoholism than did children of nonalcoholic parents. But despite this significant difference in outcome, the contribution of the genetic factor was only partial: Some children of nonalcoholic natural parents became alcoholics as adults; most children of alcoholic parents successfully avoided alcoholism.

Do statistically significant differences among groups ever have significance clinically? In our experience, such differences have clinical significance only when all or most of one group shows one pattern and few or none of another does so; in such a rare event, however, it is doubtful that one would feel the need to point to statistical confirmation for such an obvious relationship. This issue, of course, is one of the touchstones of the applied behavior analysts, most of whom have chosen to dispense with statistics in favor of research designs permitting identification of the "active ingredients" of behavior change regimens capable of modifying a given target behavior every time it (or they) is applied.

Discussion

Issues of generalization. Throughout this article, we have urged the reader to adopt a

conservative, essentially modest, approach to his/her data and to claims or conclusions deriving therefrom. For the same reasons, we wish to suggest a spirit of modesty in the Discussion section of a research paper. Two essential features of this attitude are worth emphasizing: (a) Discussion of results might well begin with a section that details limits on generalization from the data reported. If only males over the age of 40 were studied and prior data suggest that the problem under investigation is not limited to that group, it is well to point out that conclusions from the research may refer only to older males, perhaps even older males from the geographic or sociocultural group studied (if those variables might also affect generalizability). (b) Even in the absence of specific methodological constraints on generalizability, it is wisest and most sound to limit the extent to which one lays claim for the widespread or universal relevance of one's findings. A host of variables specific to one's subject sample, procedure, or data analysis, unknown or undetected during the research, could later come back to haunt one. The classic example of such an embarrassment, of course, is the unfortunate team of biochemists who reported discovery of a "schizophrenic" blood portion, only to have to report some years later that their sample of schizophrenics had consumed so much coffee that the coffee's metabolic residuals had stamped their subjects' plasma as abnormal!

Bias in discussion. Beyond overblown claims of the merits of a new therapeutic approach or unwarranted generalizations beyond sample populations, the most common problem arising from discussion of alcoholism and drug dependence data comes when the author does not discuss data in terms of prior research on both sides of his/her position.

Although this approach to science sounds unthinkable, it is surprisingly easy to adopt a position, accept it fully, then view as well done and careful only work that supports that position. To this end, one of the most cherished, widely held views by alcoholism workers has long been that the only treatment goal for alcoholics can be abstinence. This position has continued to keep its supporters, of whom there are many, from more recent

data suggesting that conventional, abstinence-oriented treatment for alcoholism is relatively ineffective (Pattison et al., 1977) and that some alcoholics do return to a controlled pattern of drinking either after treatment or in its absence (Pomerleau, Pertschuk, Adkins, & Brady, in press; Sobell & Sobell, 1976). Nonetheless, one reads startlingly ad hominem criticisms of researchers who have acknowledged the possibility that the new data on controlled drinking might justify a change in thinking. A similar lack of objectivity characterizes those who hold to the belief that alcoholism is a medical disease, wholly or largely, despite evidence to the effect that social-learning mechanisms, sociocultural influences, and psychological phenomena also play important etiologic roles in the disorder.

Reference Note

1. Homer, A. L., & Ross, S. M. *The reliability and validity of interview data from drug and alcohol abusers*. Paper presented at the National Drug Abuse Conference, San Francisco, May 1977.

References

- American Medical Association, Committee on Alcoholism. Hospitalization of patients with alcoholism. *Journal of the American Medical Association*, 1956, 162, 750.
- Anderson, G. S., & Nutter, R. W. Clients and outcomes of a methadone treatment program. *International Journal of the Addictions*, 1975, 10, 937-948.
- Amor, D. J., Polich, J. M., & Stambul, H. B. *Alcoholism and treatment*. Santa Monica, Calif.: The Rand Corporation, 1976.
- Briddell, D. W., & Nathan, P. E. Behavior assessment and modification with alcoholics: Current status and future trends. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification*. New York: Academic Press, 1976.
- Cahalan, D., & Cisin, I. H. American drinking practices: Summary of findings from a national probability sample: II. Measurement of massed versus spaced drinking. *Quarterly Journal of Studies on Alcohol*, 1968, 29, 642-656.
- Carlin, A. S., & Stauss, F. F. Descriptive and functional classifications of drug abusers. *Journal of Consulting and Clinical Psychology*, 1977, 45, 222-227.
- Chapman, L. J., & Chapman, J. P. Selection of subjects in studies of schizophrenic cognition. *Journal of Abnormal Psychology*, 1977, 86, 10-15.
- Cox, T. J., & Longwell, B. Reliability of interview data concerning current heroin use from heroin addicts on methadone. *International Journal of the Addictions*, 1974, 9, 161-165.
- Ditman, K. S. Review and evaluation of current drug therapies in alcoholism. *International Journal of Psychiatry*, 1967, 3, 248-258.
- Doleys, D. M., Ciminero, A. R., Wallach, E. S., & Davidson, R. S. Responding by alcoholics during aversive conditioning. *Behavior Modification*, 1977, 1, 205-220.
- Goldberg, L. R. Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 1968, 23, 483-496.
- Goodwin, D. W., & Guze, S. B. Heredity and alcoholism. In B. Kissin & H. Begleiter (Eds.), *The biology of alcoholism* (Vol. 3). New York: Plenum Press, 1974.
- Gotestam, K. G., Melin, L., & Öst, L.-G. Behavioral techniques in the treatment of drug abuse: An evaluative review. *Addictive Behaviors*, 1976, 1, 205-225.
- Hedberg, A. G., & Campbell, L. A comparison of four behavioral treatments of alcoholism. *Journal of Behavior Therapy and Experimental Psychiatry*, 1974, 5, 251-256.
- Keller, M. The oddities of alcoholics. *Quarterly Journal of Studies on Alcohol*, 1972, 33, 1147-1148.
- Levine, J., & Zigler, E. Humor responses of high and low premorbid competence alcoholic and nonalcoholic patients. *Addictive Behaviors*, 1976, 1, 139-150.
- Marlatt, G. A. The drinking profile: A questionnaire for the behavioral assessment of alcoholism. In E. J. Mash & L. G. Terdal (Eds.), *Behavior therapy assessment: Diagnosis and evaluation*. New York: Springer, 1975.
- Marlatt, G. A., Demming, B., & Reid, J. B. Loss of control drinking in alcoholics: An experimental analogue. *Journal of Abnormal Psychology*, 1973, 81, 233-241.
- McCabe, O. L., Kurland, A. A., & Sullivan, O. Paroled narcotic addicts in a verified abstinence program: Results of a five-year study. *International Journal of the Addictions*, 1975, 10, 211-228.
- Miller, P. M. A behavioral intervention program for chronic drunkenness offenders. *Archives of General Psychiatry*, 1975, 32, 915-918.
- Miller, P. M. *Behavioral treatment of alcoholism*. New York: Pergamon Press, 1976.
- Miller, P. M., & Eisler, R. M. Alcohol and drug abuse. In W. E. Craighead, A. E. Kazdin, & M. J. Mahoney (Eds.), *Behavior modification principles, issues, and applications*. Boston: Houghton Mifflin, 1975.
- Mirin, S. M., Meyer, R. E., McNamee, H. B., & McDougale, M. Psychopathology, craving, and mood during heroin acquisition: An experimental study. *International Journal of the Addictions*, 1976, 11, 525-544.
- Nathan, P. E. *Cues, decisions, and diagnoses*. New York: Academic Press, 1967.
- Nathan, P. E., & Briddell, D. W. Behavioral assessment and treatment of alcoholism. In B. Kissin

- & H. Begleiter (Eds.), *The biology of alcoholism* (Vol. 5). New York: Plenum Press, 1977.
- Nathan, P. E., & Lansky, D. Management of the chronic alcoholic. In J. P. Brady & H. K. Brodie (Eds.), *Controversy in psychiatry*. Philadelphia, Penn.: Saunders, 1978.
- O'Leary, M. R., Radford, L. M., Chaney, E. F., & Schau, E. J. Assessment of cognitive recovery in alcoholics by use of the Trail-Making Test. *Journal of Clinical Psychology*, 1977, 33, 579-582.
- Pattison, E. M., Sobell, M. B., & Sobell, L. C. (Eds.). *Emerging concepts of alcohol dependence*. New York: Springer, 1977.
- Pomerleau, O. F., Pertschuk, M., Adkins, D., & Brady, J. P. A comparison of behavioral and traditional treatment for middle income problem drinkers. *Behavior Therapy*, in press.
- Rawlins, M., Randall, M., Meyer, R. E., McNamee, H. B., & Mirin, S. M. Aftercare on narcotic antagonists: Prospects and problems. *International Journal of the Addictions*, 1976, 11, 501-512.
- Rosenthal, R. *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts, 1966.
- Schaefer, H. H., Sobell, M. B., & Mills, K. C. Baseline drinking behavior in alcoholics and social drinkers: Kinds of drinks and sip magnitude. *Behaviour Research and Therapy*, 1971, 9, 23-27.
- Selzer, M. L., Vinokur, A., & Wilson, T. D. A psychosocial comparison of drunken drivers and alcoholics. *Journal of Studies on Alcohol*, 1977, 38, 1294-1312.
- Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 1977, 32, 752-760.
- Sobell, L. C. A critique of alcoholism treatment evaluation. In G. A. Marlatt & P. E. Nathan (Eds.), *Behavioral assessment and treatment of alcoholism*. New Brunswick, N.J.: Rutgers University Center of Alcohol Studies, 1978.
- Sobell, M. B., & Sobell, L. C. Alcoholics treated by individualized behavior therapy: One-year treatment outcome. *Behaviour Research and Therapy*, 1973, 11, 599-618.
- Sobell, M. B., & Sobell, L. C. Second-year treatment outcome of alcoholics treated by individualized behavior therapy. *Behaviour Research and Therapy*, 1976, 14, 279-288.
- Vogler, R. E., Compton, J. V., & Weissbach, T. A. Integrated behavior change techniques for alcoholics. *Journal of Consulting and Clinical Psychology*, 1975, 43, 233-243.
- Wilson, G. T., Leaf, R., & Nathan, P. E. The aversive control of excessive drinking by chronic alcoholics in the laboratory setting. *Journal of Applied Behavior Analysis*, 1975, 8, 13-26.

Received September 29, 1977 ■

Methodological Issues in Research with Correctional Populations

N. Dickon Reppucci and W. Glenn Clingempeel
University of Virginia

Methodological issues confronting psychologists who conduct research with correctional populations are examined. The article is divided into five major sections: (a) the intrusion of values; (b) problems with trait-derived methodologies in correctional population research, including inadequate construct validity, ignoring environmental influences and ignoring subjects; (c) naturalness and the problem of external validity; (d) recidivism; and (e) ethics. General recommendations are offered at the conclusion of each section.

The purpose of the present article is to discuss "methodological issues" confronting psychologists who conduct research with correctional populations. In the selection of issues for discussion—a necessarily value-laden process—we have chosen to focus on problems emanating from ignoring external validity considerations (e.g., In what ways can the findings be generalized?) and from the potentially constrictive impact of prevailing values and conceptual weltanschauung on research designs. The reader should not interpret this focus as indicating that internal validity (e.g., Did the experimental condition make a difference?) and/or data analysis issues (e.g., the use of univariate analyses when multivariate analyses are appropriate) are either unimportant or uncommon in correctional research. Such an interpretation would be erroneous. Rather, the decision to focus on the former issues, although not entirely excluding the latter, is based on the assumption that these latter problems are generally given center stage in methodological critiques of psychological research and that another similar critique would be less informative than the current attempt. The accuracy of this judgment is left to the reader.

In attempting to delimit this critique, sev-

eral constraints and qualifications should be underscored:

1. The problems discussed are not necessarily unique to research with correctional populations. Many, if not all, issues could just as easily be raised with regard to other involuntarily incarcerated populations, for example, mental patients. Indeed, some of the issues are salient for much of psychological research, in general, and for personality research with "clinical" populations, in particular.

2. Since there are recent excellent reviews of methodological problems of behavior modification and delinquency studies (e.g., Davidson & Seidman, 1974; Emery & Marholin, 1977) and of large-scale program evaluation research in correctional settings (e.g., Adams, 1974; Glaser, 1974; Sarri & Selo, 1974), these areas are not considered.

3. Only research actually conducted with subjects residing in correctional settings was examined. This constraint allowed exclusion of those studies focusing on experimental interventions with offenders subsequent to their release from a correctional facility.

4. The data for deriving and elucidating the issues discussed in this article are in no way all inclusive, being largely based on only two sources: (a) a scrutiny of unpublished articles reviewed by the first author over the past 5 years in his role as a consulting editor for the *Journal of Consulting and Clinical Psychology (JCCP)* and (b) an intensive

Requests for reprints should be sent to N. Dickon Reppucci, Department of Psychology, Gilmer Hall, University of Virginia, Charlottesville, Virginia 22901.

Table 1

Studies Using Correctional Populations Published in the Journal of Consulting and Clinical Psychology (JCCP) and the Journal of Abnormal Psychology (JAP) During the Period January 1967–August 1977

Type	JCCP	JAP
Personality discrimination		
Correctional subgroup model ^a	Blackburn, 1969; Borkovec, 1970; Erikson & Roberts, 1971; Ganzer & Sarason, 1973; McGuire & Megargee, 1974; Roberts, Erikson, Riddle, & Bacon, 1974; Moss, Hosford, Anderson, & Petracca, 1977	Hare, 1968; Mosher, Mortimer, & Grebel, 1968; Oliver & Mosher, 1968; Cohen, Seghorn, & Calmas, 1969; Skrzypek, 1969; Roberts & Erikson, 1969; Orris, 1969; Schmid, 1970; Stewart, 1972; Kercher & Walker, 1973; Davids & Falkof, 1975
Noncorrectional control model ^b	Bixenstine & Buterbaugh, 1967; Stein, Sarbin, & Kulik, 1968; Kulik, Stein, & Sarbin, 1968a; Kulik, Stein, & Sarbin, 1968b; Landau, 1976	Stewart & Resnick, 1970; Schlichter & Ratliff, 1971; Sutker, 1971; Hetherington, Stouwie, & Ridberg, 1971; Jurkovic & Prentice, 1977
Treatment evaluation	Schwitzgebel, 1969; Sowles & Gil, 1970; Jeansess, 1975	Truax, Wargo, & Volksdorf, 1970; Redfering, 1973
Adjustment prediction	Cowden & Pacht, 1967	
Test construction		Megargee, Cook, & Mendelsohn, 1967; Carlson, 1972
Miscellaneous ^c	Armentrout & Rouzer, 1970; Grisso, 1975; McCandless, Persons, & Roberts, 1972; Gunn & Gristwood, 1975	Roberts & Erikson, 1968; Ruma & Mosher, 1967; Persons & Marks, 1970; Rotenberg & Sarbin, 1971
Total	20	24

^a Subcategories of correctional populations (e.g., psychopathic and neurotic delinquents) are compared.

^b A correctional population is compared with a noncorrectional population.

^c Contains all studies that could not be unequivocally classified in the other four categories.

examination of all relevant studies published in *JCCP* and the *Journal of Abnormal Psychology* (*JAP*) during the last 10½ years (January 1967–August, 1977) (see Table 1).¹

5. Attention is focused on what we call the *personality discrimination* research paradigm, since, as is clear from Table 1, over 60% of the published studies examined were subsumed under this category. (The majority of unpublished studies submitted to *JCCP* also fell in this category.) In brief, this paradigm uses population subtypes (e.g., delinquents and nondelinquents, recidivists and nonrecidivists) as the independent measures and attempts to determine from an array of dependent measures (e.g., measures of impulsivity, delay of gratification, sensation

seeking, time perspective) those that discriminate between the groups.

This article is divided into five major sections. The first section examines the intrusion of certain value assumptions on the research enterprise and the methodological stagnation that we believe has resulted from these intrusions. The second section confronts three shortcomings fostered by trait-derived methodologies, including the problem of inadequate

¹ These two journals were selected for intensive examination because they are the two APA-sponsored journals most likely to publish research articles containing correctional populations and therefore of most potential interest to readers of this journal.

construct validity, the problem of ignoring environmental influences, and the problem of ignoring subjects. The third section focuses on external validity problems stemming from the failure to incorporate naturalistic dimensions in research designs. The fourth section outlines some of the difficulties with the much used recidivism measure. The final section provides a few comments on the ethics of research with correctional populations. Each section is concluded with one or more recommendations.

The Intrusion of Values

To suggest that values pervade every stage of the research enterprise from choice of problem and selection of research design to analysis and interpretation of data is only to repeat what numerous others have said before (e.g., Allport, 1940, 1943; Hudson, 1972; Sarason, 1975). Yet the failure to confront the potential impact of implicit value assumptions on the research process is a ubiquitous shortcoming endemic to much of social science research (Caplan & Nelson, 1973). In this section, attention is focused on a few specific conceptual-methodological constrictions engendered by values pervading research with correctional populations.

From the review of studies with correctional subjects (see Table 1), two major value premises emerged—(a) the assumption of offender deficit and (b) the assumption of discriminating traits. Although not exhaustive of value influences, these two assumptions permeated almost all of the research that was examined.

The assumption of offender deficit is based on the belief that something must be psychologically wrong with any person who has committed a crime (Brodsky, 1973). This assumption appears to have contributed to much research whose goal is the search for psychological or personality disorder in prisoners. As a result there is an overemphasis on disorder, psychopathology, and other negative characteristics. Unfortunately, this emphasis characterizes much behavioral science research with minority and/or "clinical" groups (Caplan & Nelson, 1973; Jourard, 1968; West & Gunn, Note 1). Recently, Caplan and Nelson

(1973) demonstrated this preoccupation in psychological research with black Americans. They examined all studies of black Americans appearing in the first 6 months of the 1970 *Psychological Abstracts*. In 82% of the studies that they classified, the nature of the research suggested personal deficits as the cause of difficulties for black Americans. They concluded that

the picture that emerges is one of psychologists investing disproportionate amounts of time, funds, and energy in studies that lend themselves, directly or by implication, to interpreting the difficulties of black Americans in terms of personal shortcomings. (Caplan & Nelson, 1973, p. 204)

The major point to be drawn from their analysis is the analogous one that the offender deficit assumption constricts the research questions asked and the methods used. Two major omissions, for example, of much of the psychological research on criminal behavior are (a) focusing on prisoner strengths, for example, studies of helping behavior or other positive characteristics, and (b) ignoring the potential impact of situational and environmental factors in precipitating the criminal act. These omissions are viewed as a by-product of the attribution of personal deficits. If the causes of delinquency are defined in negatively valenced, person-centered terms (e.g., impulsivity, inability to delay gratification), then the inclusion of positively valenced terms and situational or environmental factors in the research design is neglected.

The second pervading value premise, derived from nomothetic personality theory, is the assumption of discriminating traits. This assumption suggests that a set of trait dimensions is applicable to all persons and that individual differences are to be identified with different locations on these dimensions. Moreover, on specific traits (e.g., impulsivity, aggressiveness), correctional subjects are presumed to be distant enough on the same trait continuum to enable discrimination from non-correctional subjects. A corollary assumption is that these discriminating characteristics are relatively enduring and cross-situationally consistent. It is implied that the rank order of individuals on any trait should remain the same in all situations. This assumption has given rise to methods (as in the personality

discrimination model) that ignore person-situation interactions and the advantages of idiographic strategies for personality assessment and prediction. Elaboration of these omissions is discussed elsewhere in this article (see *The Problems of Ignoring Environments and Ignoring Subjects*). At this point, it is necessary only to emphasize that in our opinion, conceptions of the personality of criminal offenders as internal behavior dispositions have resulted in a constriction of methods, a correspondingly reduced predictive utility, and incomplete conceptions of criminal behavior.

Recommendation

Investigators should recognize that the value assumptions of offender deficit and discriminating traits underlie much of the psychological research with correctional populations. Future research should focus on strengths as well as weaknesses and the development of experimental designs that do not neglect environmental influences and person-situation interactions.

Problems with Trait-Derived Methodologies in Correctional Population Research

Recently, Endler and Magnusson (1976) have argued convincingly that personality theory differences are inextricably linked to differences in research strategies and measurement models. For example, research stemming from trait model assumptions includes methods (e.g., questionnaires, ratings, tests) that are often different from those methods stemming from operant-type situationism assumptions (e.g., behavioral response rates and behavioral checklists for functional analysis of individual behavior). This linkage of personality theory and methodology is crucial to our analysis.

As pointed out in the previous section, trait assumptions and concomitant methodologies have occupied a prominent role in personality research with correctional populations. The frequency of use of the personality discrimination model—with its potentially discriminating measures of traits obtained in similar situations through similar response

modes—is a testimony to the currently functional status of this preoccupation. Despite its popularity in correctional research, the assumptions of the trait model have been frequently assailed. For example, an abundance of empirical evidence tending to refute the cross-situational consistency assumption has emerged from three sources (Endler & Magnusson, 1976)—(a) research (partitioning the variance into person, situation, and person-situation interaction components) that demonstrates that person-situation interactions account for more of the behavioral variance than persons alone and situations alone (Bowers, 1973; Endler & Hunt, 1966; Endler & Magnusson, 1976; Mariotti & Paul, 1975; Moos, 1968, 1969; Rausch, Farbman, & Llewellyn, 1960); (b) correlational research strategies in which correlations of same traits in different situations are found to be moderately low (Hartshorne & May, 1928; Magnusson, Heffler, & Nyman, 1968; Rushton, 1976); and (c) studies (simultaneously including personality and situational variables as independent measures in the design) that reveal interaction effects rather than main effects as the true state of affairs (Baron, Cowan, Ganz, & McDonald, 1974; Cronbach & Snow, 1975). From a review of these and other relevant studies, various investigators have concluded that there is a paucity of empirical support for the assumption of cross-situational consistencies of behavior (Argyle & Little, 1972; Endler, 1973, 1975a, 1975b; Endler & Magnusson, 1976; Mischel, 1973, 1977). With the acceptance of this conclusion, the task in this section is to focus specifically on the limitations of trait-derived methodologies as they have been used in personality research with correctional populations. Three of these problems, construct validity, ignoring environments, and ignoring subjects, are examined in the following discussion.

The Problem of Construct Validity

One salient problem that emerges from the trait-derived personality discrimination studies is the lack of attention given to construct validity issues (Cronbach & Meehl, 1955). Messick (1975) defined construct validity as

the process of marshaling evidence in the form of theoretically relevant empirical relations to support the inference that an observed response consistency has a particular meaning" (p. 955). Campbell (1960) outlined two requirements of construct validity, which he labeled *trait* and *nomological* validity. Trait validity refers to the requirement that the domain of observables that purportedly measure a particular construct must demonstrate internal consistency when put to the empirical test. This means that measures of the same construct must correlate highly with one another in individual difference studies and/or be similarly affected by experimental treatments. Nomological validity refers to the requirement that the measures of one construct behave in accordance with the network of relations to other constructs as derived from a formal theoretical system.

The trait validity requirement suggests that two or more measures of a personality dimension (e.g., impulsivity) should be substantially interrelated. For example, if institutionalized delinquents were to exhibit more impulsivity than nondelinquents on the Barratt Impulsivity Scale (Barratt, 1959), then evidence for construct validity would be obtained if this greater impulsivity also manifested itself on a second measure (e.g., the Hirschfield Impulsivity Scale; Hirschfield, 1965). In addition, both measures of impulsivity should behave similarly with regard to discriminating a delinquent from a nondelinquent sample. Construct validity could be questioned if the delinquent sample scored significantly different from the nondelinquent sample on the Barratt Impulsivity Scale but no differences or the reverse finding emerged on the Hirschfield Scale.

Although it is important that measures of the same construct exhibit a convergence and behave similarly, Campbell and Fiske (1959), in a classic, albeit often ignored, article, maintained that construct validation requires evidence of discriminant as well as convergent validity. Evidence for discriminant validity is revealed when measures of one construct are shown to have weak interrelations with measures of distinctly different constructs. Adopting the discriminant validity requirement, evidence for construct validity would

be weakened if personality measures of one dimension (e.g., impulsivity) correlated too highly with personality measures of another (e.g., nurturance). In the articles examined, the concept of discriminant validity was never mentioned. Clearly, a problem with the research that seeks to determine the extent to which correctional subjects differ from non-correctional subjects on some construct dimension is the failure to use multiple measures of both the focal construct and different constructs. As such, there tends to be little evidence for either convergent or discriminant validity in most studies of correctional populations.

A second construct validity problem of the personality discrimination studies hinges on the failure to take into account that each measure of a construct represents a trait *and* a method unit. The fact that two measures of impulsivity are interrelated may reflect only that they were measured by the same method (e.g., self-report) and not that they "tap" the same construct. Moreover, the reliance on such methods as self-report questionnaires and rating scales in the personality discrimination studies suggests the possibility that the observed consistencies may be more in response to method constraints than true differences on the construct. It is possible, for example, that an "oppositional" or "deviation" response set in which correctional populations choose responses that they believe differ from expectations could account for consistent performances on self-report questionnaires. Campbell and Fiske (1959) proposed the multitrait-multimethod matrix as a solution to both the convergent-discriminant validity problem and the problem of method-trait confounds. In this triangulation method, two or more traits are correlated with two or more methods. This procedure would suggest, for example, that the two traits, impulsivity and hostility, be measured by at least two methods (e.g., self-report and behavioral observations). Evidence for convergent validity would be obtained if the correlation of the same trait (e.g., impulsivity) by different methods (e.g., self-report and behavioral observation) was high. Evidence for discriminant validity would be obtained if (a) the same trait/different method (e.g., impulsivity

by self-report with impulsivity by behavioral observations) correlations were higher than different trait/different method (e.g., impulsivity by self-report with nurturance by behavioral observations) correlations, and, most importantly, (b) the same trait/different method correlations should also be higher than different trait/same method (e.g., impulsivity by self-report with nurturance by self-report) correlations. In effect, evidence for construct validity is obtained when measures with only trait variance in common exhibit higher correlations with each other than measures with common method variance and measures with neither trait nor method variance in common.

In summary, multitrait-multimethod techniques have seldom been used in research with correctional populations. As a result the construct validity of those measures in current use for discriminating correctional from noncorrectional populations remains uncertain at best. Although recognizing that evidence for construct validity is obtained when measures of a particular construct consistently discriminate between population groups presumed different on that construct, we contend that certain facets of the construct validation process have been paid relatively little attention. It is perhaps worth noting that in one study (Saunders, Reppucci, & Sarata, 1973) that attempted to marshal evidence for the construct validation of impulsivity as a trait characterizing delinquents, the investigators concluded that there was no evidence for the hypothesis that delinquent behavior is related to impulsivity or that existing measures of impulsivity tap the same dimension. Although the process of construct validation is a time-consuming venture, requiring multiple studies and the gradual weaving of a net of circumstantial evidence around a construct, its importance should not be underestimated. Personality labels that are supposedly descriptive of inherent traits of correctional populations often contribute to decisions regarding incarceration and to the development of intervention programs.

Recommendation

Increased attention in personality research with correctional populations should be given

to the construct validation process. As a minimum requirement, research designs should include multiple measures of each attribute or construct under investigation with those measures reflecting at least two different methods. Correlations among measures should be obtained.

The Problem of Ignoring Environments

The main criticism of the trait model has been its empirical neglect of situational factors as they affect an individuals' behavior (Endler & Magnusson, 1976). Consistent with this contention, the frequent use of trait-derived methodologies in correctional population research has resulted in the ignoring of possible interactive influences of environmental and situational factors in research designs. Significant predictive and generalizability limitations have been the consequence. This "problem of ignoring environments" has manifested itself in at least three forms in the personality discrimination studies and one form in the adjustment prediction studies.

First, with an assumption of cross-situational consistency ostensibly prevailing, personality discrimination studies have monitored a narrow universe of response modes in a narrow universe of situations. In a typical study, for example, subjects consisting of either matched groups of correctional population subtypes (e.g., neurotic and psychopathic delinquents) or a correctional group and a noninstitutional nonoffender control (e.g., high school students) are administered individually or in small groups an array of self-report questionnaires. Any differences that emerge between the groups are viewed as relatively stable and situation-free personality characteristics. The possible differential influences of situational factors are not considered. This holding of situations constant within and between studies has sustained a bias toward proving the trait notion (therefore increasing the likelihood of positive results).

Although incarcerated delinquents may, in fact, prefer immediate over delayed gratification more so than nonincarcerated nondelinquents as assessed by responses to hypotheti-

cal questions, the relative position of these groups on the delay of gratification continuum may shift as situational factors change (e.g., as the delinquents are released). In fact, Mischel (1973) has pointed out that to predict a subject's delay of gratification behavior, one may need to know such moderator variables as how old he/she is, the experimenter's sex, the particular objects for which the subject is waiting, and the consequences of not waiting. Given the possible influences of similar situational variables and their potential interaction with personality variables, the predictive utility or generalizability of good discriminators may be significantly impaired.² What is neglected in trait-derived methodologies is the attempt to answer the question, How are persons who end up in prisons different from each other and from noncorrectional populations with regard to specific personality characteristics as measured across specific situations? Clearly, then, to determine Personality \times Situation interactions, there is a need to incorporate both situational and personality variables as independent measures in research designs.

A second form of ignoring environments centers on the inadequate description of the experimental situation itself. The typical personality discrimination study cursorily mentions that an experimenter (whose characteristics are not described) administered a series of questionnaires (or tasks) in one or more settings (which are also usually not described) to two or more groups of subjects. Both internal and external validity problems may result. The internal validity problem is that the microsetting of the experiment, including the personal characteristics of the experimenter, may interact differently with the groups and significantly influence any observed group differences. The external validity problem is that inadequate descriptions render generalizability across experimenters, situations, and tasks a most hazardous business.

One area of inadequate description focuses on the experimenter. Although various experimenter characteristics have been shown to have demonstrable effects on the behavior of subjects in experimental situations (e.g., Barber & Silver, 1968; Rosenthal, 1963, 1968), McGuigan's (1963) characterization of the

experimenter as "the neglected stimulus object" certainly holds for the studies reviewed for this article. The fact that such basic experimenter characteristics as sex, age, and attractiveness may differentially affect the performance of incarcerated versus nonincarcerated males and/or females is simply not taken into account. In only 1 of the 28 published personality discrimination studies was there any description of the experimenter. This prevailed despite the fact that in 19 of these studies, subjects were evaluated individually—a situation that permits potentially greater experimenter impact. Moreover, although 18 of these studies had multiple authors, which presumably means a high likelihood of multiple experimenters (McGuigan, 1963), no studies attempted to investigate experimenter characteristics as an independent variable. This state of affairs seems increasingly dubious in view of a recent wave of investigations demonstrating that different experimenters obtain significantly different results with different subjects (Rumenik, Capasso, & Hendrick, 1977; Silverman, 1974; Silverman, Shulman, & Wiesenthal, 1972).

Neglecting the effect of experimenter characteristics engenders an internal validity problem; namely, the same or different experimenters may affect the performance of subjects in different groups in different ways. Stewart and Resnick's (1970) verbal conditioning study illustrates how this difficulty might manifest itself. These authors compared 33 incarcerated delinquent males with a control group of 33 boys from a nearby high school on a task measuring conditionability of aggressive and dependency verbs. All subjects were tested individually by an experimenter (or experimenters) who was (were) not described. In accordance with predictions, the delinquent group rejected the dependency verbs more so than did the nondelinquent group. The authors concluded that delin-

² A particular measure may reveal a highly significant difference between groups (and thus be a good discriminator) and yet be a poor predictor. Therefore, the discrimination-to-prediction leap should only occur after empirical investigations have confirmed the predictive utility of good discriminators.

quents found it more difficult to express dependency behavior than nondelinquents. Although not necessarily an unreasonable conclusion given a recognition of generalizability limitations, it is nonetheless quite possible that undescribed experimenter(s) characteristics were a (or the) major contribution to performance differences between the two groups. For example, depending on the experimenter's age, sex, and perceived status, he or she may have engendered less (or more) cooperation from the incarcerated delinquent group than from the high school group. Thus, differences between the groups may have resulted from the type of experimenter(s) used rather than because of personality differences between the boys. The point is that we do not know, because no information regarding the effect of experimenter characteristics was provided.

Experimenter neglect also constitutes an external validity, generalizability of findings, problem. Levine (1974) captured the crux of this concern:

The generalization possible from any experiment is limited to the set of conditions that includes a definition of that universe of experimenters and their behavior, of which the experimenters used in the experiment are a sample taken according to a well-defined principle of sampling. (p. 663)

Beyond the ignoring of possible experimenter effects, there tends to be both inadequate observation and description of the microculture of the experimental situation. The physical characteristics of the testing situation, including the behaviors of experimenter and subject that are seemingly peripheral to the study, and the nature of the rapport-establishing process, may singly or in combination exert subtle influences. Yet, in only a few of the studies examined were these dimensions adequately described. While recognizing that factors of this nature have allegedly been held constant and thus may not have significantly affected the results of the published personality discrimination studies, we also recognize that negative results and inconsistencies have a habit of not getting published. Moreover, the possible influences of higher order interactions become more important as one ventures beyond same situation/same response mode investigations to

develop studies that compare different subject types across different situations and that use more naturalistic, nonreactive measures. While realizing the impossibility of incorporating a multitude of potential moderator variables into any single experimental design, we suggest that a more ethological perspective—with a concomitant sharpening of observational and descriptive skills—may be useful in the understanding of inconsistencies and the possible influences of higher order interaction effects.

The primary concern is that researchers avoid misinterpreting Morgan's canon. As Cronbach (1975) stated:

From Occam to Lloyd Morgan, the canon has referred to parsimony in theorizing, not in observing. The theorist performs a dramatists' function; if a plot with a few characters will tell the story, it is more satisfying than one with a crowded stage. But the observer should be a journalist, not a dramatist. To suppress a variation that might not recur is bad observing. (p. 124)

The ignoring of the impact of correctional institutions on the behavior of offenders signifies a third form of environmental neglect. This neglect spawns two kinds of problems for personality research with correctional populations. The first problem centers on the failure in studies which search for discriminating personality characteristics of criminal offenders to extricate the effects of institutionalization from the effects attributable to offender characteristics. The subset of personality discrimination studies that compares institutional and noninstitutional populations illustrates this problem. In only 1 (Landau, 1976) of the 10 published studies was there any attempt to control for the impact of institutionalization (independent of subject characteristics) on performance of the dependent measures.

Thus, differences that have emerged between correctional and noncorrectional populations may be more a function of the impact of institutionalization than of differences between the groups on particular personality characteristics. Certainly commentaries on the possible profound effects of institutional environments on behavior (e.g., Goffman, 1961; Zimbardo, 1973) alert us to this likelihood. Landau (1976), in a study of time ori-

entation, recently attempted to remedy this problem. He included subject groups of institutionalized delinquents, noninstitutionalized delinquents (delinquents on probation), institutionalized nondelinquents (soldiers), and noninstitutionalized nondelinquents (students in a vocational training school). He also controlled for length of institutionalization for the incarcerated delinquents and the soldiers. The results provided strong evidence for the independent effect of institutionalization on various aspects of past, present, and future time perspective. Although (as might be expected) the army proved to be a milder form of institutionalization than prison, the effect of institutionalization was nevertheless substantial. Landau (1976) concluded that

this study provides additional evidence of the strong and measurable effect of institutionalization on various aspects of time orientation. This should be considered as a significant example of the possible effect that salient situational factors may have on characteristics and behaviors that are too frequently considered as stable and situation-free personality traits. (p. 757)

The second problem associated with the ignoring of institutionalization effects is the problem of inducing comparability of research findings conducted in different correctional institutions. The problem is accentuated by empirical evidence suggesting that correctional institutions differ on a number of dimensions (e.g., the treatment-security continuum) and that these differences are reflected in their effects on inmates (Moos, 1975; Street, Vinter, & Perrow, 1966). For example, Street et al.'s (1966) comparative evaluation of six correctional institutions revealed that institutional differences in treatment goals and philosophies were significantly related to inmate attitudes toward the total program, staff, other inmates, and themselves. Institutions with different orientations (e.g., control and obedience, reeducation and development, or treatment) clearly had different effects on their inmates. Moreover, although some of the variance could be accounted for by characteristics of the inmate populations, the independent effects of the organizations and their different orientations were pronounced. Additional evidence of correctional institution diversity comes from Moos's

(1975) cluster analytic derivation of six distinct institution types vis-à-vis the Correctional Institutions Environmental Scale (CIES). The major point is that regardless of whether the CIES, structural-organizational assessments, and/or alternative institutional description methods are used, institutional description strategies of some form are required if researchers earnestly seek comparability of research findings obtained in different correctional institutions.

Furthermore, it should be emphasized that correctional institutions are not homogeneous settings but contain a diversity of subsettings or situations, some of which are common across institutions, whereas others may be unique to a specific institution. Therefore, future researchers should elucidate and dimensionalize the crucial situations within and between correctional institutions. With regard to methods, Magnusson and Ekehammar's (1975) suggestion that we study *situation perception* (the meaning that an individual assigns to a situation), *situation reaction* (based on individual's responses to situations), and the relationship between these dimensions seems particularly worthy of adherence. The Person \times Situation interaction studies could then take the form of more naturalistic studies of person subtypes across actual situations within the institution. An important caveat is in order. The study of Person \times Situation interactions within correctional institutions will not likely give us much information about Person \times Situation interactions outside of the institutional setting. Obviously, relative to the extrainstitutional environment, in prisons the number of available situations as well as the freedom to choose among and capacity to change such situations is markedly delimited. Thus, there should be no expectation that research with incarcerated offenders will shed much light on the question of which specific Person \times Situation interactions outside the prison will be likely to result in the commission of which specific law-violative behaviors.

The major form of ignoring the environment in the adjustment prediction studies³

³ Adjustment prediction studies contain an array of personality, demographic, and institutional per-

has been the lack of attention paid to characteristics of the postinstitutional environment to which the correctional subject returns. Typically, the array of predictors of recidivism and postinstitutional adjustment have consisted of demographic, personality, and institutional adjustment measures. The availability of formal support systems (e.g., jobs, recreational facilities, financial resources, educational opportunities) and informal social networks (e.g., family, extended kin, friends), both of which may have a significant impact on the process of postinstitutional adjustment (see, e.g., McArthur, 1974), are seldom included in the prediction equation. Recently, Mischel (1977) has argued in a corresponding situation with mental patients that "accurate predictions of posthospital adjustment require knowledge of the environment in which the ex-patient will be living . . . rather than any measured person variables or in-hospital behavior" (p. 251).

Perhaps the best example of the inability of the typical psychological and demographic data to predict the future behavior of prisoners and others is found in the studies of the prediction of violent behavior (Megargee, 1970; Monahan, 1976; Wenk, Robison, & Smith, 1972). Not only has univariate analysis proved woefully inadequate, but there is also evidence to suggest that combining a multitude of person-centered predictors using sophisticated multivariate analyses is insufficient (Monahan, 1976). Most studies indicate a false positive prediction rate of over 80%, and the most accurate prediction of true positives (46%) still mispredicted 54% of the subjects (Monahan, 1976). In addition, Heller and Monahan (1977) pointed out that "the studies that report the highest percentage of true positives are generally those with the weakest methodologies" (p. 137). With the prediction difficulties encumbering person-centered approaches to violence prediction, recommendations have surfaced that ecological evaluations (e.g., predicting which situa-

tions will elicit violent behavior) replace the more common individual-oriented strategies.

Recommendations

1. Psychologists should adopt a cautious approach to inferring behavioral consistency based on responses to a specific task under a specific set of experimental conditions. Adequate descriptions of the experimental situation should be provided. When possible, multiple situations should be incorporated into the research design as independent variables. Data analysis should account for Person \times Situation interactions.

2. Adequate descriptions of the experimenter(s) should be provided. When possible, multiple experimenters should be incorporated into the research design as independent variables.

3. The factor of institutionalization should be controlled for in the experimental design. In addition, institutional description strategies (e.g., structural-organizational assessments) should be used to facilitate comparisons of correctional institutions and their differential effects. Future research should begin to dimensionalize the important situational variables within and between correctional institutions and to incorporate these variables in Person \times Situation studies. Finally, psychologists should proceed with extreme caution, if at all, in drawing conclusions about extrainstitutional Person \times Situation interactions from interaction studies conducted within the correctional environment.

4. Ecological evaluations that incorporate measures of the postinstitutional environment of prisoners, such as existing support systems and social networks, should become a standard part of adjustment prediction studies.

The Problem of Ignoring Subjects

A third pervasive problem of trait-derived methodologies is the problem of ignoring subjects. Partially as a result of the preoccupation with the nomothetic assumption that a set of trait dimensions is universally applicable and that all individuals can be identified with different locations of these dimensions, the correctional subject (as most others) has

formance variables in an attempt to determine the best predictors (through multiple regression analysis) of such criterion variables as postrelease adjustment and recidivism rates.

been neglected as an expert on his/her own behavior.

In not one of the studies listed in Table 1 were the subjects asked about their perceptions of, attitudes toward, or behavior in the research in which they were participants. Yet, subjects, in most cases are not passive recipients, even if cooperative; rather, they will develop hypotheses about the nature and purposes of the investigation. These hypotheses may be interpreted as demand characteristics that influence the subject's behavior and responses.

One potentially problematic demand characteristic is the general attitude of the correctional subject toward research and the concomitant motivation for participation. In comparison with free-world citizens, the attitudes of prisoners toward participation in research has been overwhelmingly positive (Wilson & Donnerstein, 1976). Correctional subjects may believe that participation in research is an indication of "good behavior" and therefore linked to getting out. Another possibility is that prisoners are motivated to participate in research by their boredom and the generally negative aspects of prison living (Brodsky, Note 2).

A second potential demand characteristic problem has been referred to as multiple treatment inference (Campbell & Stanley, 1963). The essence of this difficulty is that the effect of prior treatments and research participation is not erasable. This refers to multiple treatments in the same experiment as well as to research experience in other experiments. Brody (Note 3) provided an example of this problem. In an attempt to manipulate the choice of a delayed and immediate reward for an institutionalized delinquent and a noninstitutionalized "normal" sample, the delinquents, in contrast to previous research findings, consistently chose the delayed reward across all experimental conditions. After investigation of his subjects, Brody attributed this finding to the fact that similar research had been carried out in that institution only recently and that the subjects were no longer naive to the manipulations.

It may well be that the best way to alleviate the possible effect of these and other demand characteristics would be to enlist the

subject as an expert on his/her own behavior, both in the experimental situation and in general. Recently, Bem and Allen (1974) demonstrated the wisdom of enlisting the subject's self-knowledge to increase predictive power. They hypothesized that individuals who identified themselves as consistent on a particular trait dimension would behave more consistently cross-situationally than those who identified themselves as highly variable. Their results supported the hypothesis demonstrating consistency for some of the people some of the time. They then argued that a shift to idiographic rather than nomothetic assumptions about individual differences would improve predictive utility and that more attention should be paid to both persons and situations. This shift is supported by Jones and Nisbet (1971), who demonstrated that when explaining the behavior of others, people tend to invoke consistent personality dispositions, but when explaining their own behavior, they consider specific situations.

In summary, we agree with Mischel (1977) in arguing for a move toward more idiographic functional analyses:

The moral, for me, is that it would be wise to allow our "subjects" to slip out of their roles as passive "assesseees" or "testees" and to enroll them, at least sometimes, as active colleagues who are the best experts on themselves and are eminently qualified to participate in the development of descriptions and predictions—not to mention decisions—about themselves. (p. 249)

However, we also agree with Allport (1937) that idiographic and nomothetic methods are "overlapping and contributing to one another" (p. 22). Thus it would be unproductive to scuttle one approach in favor of the other.

Recommendation

Psychologists should take the role of the subject as expert seriously, since subjects know more than anyone else about situational influences on their own behavioral consistency. At the minimum, a postexperimental inquiry should be used as a standard experimental procedure to gather data on the operation of demand characteristics; at the maximum, the subject should be enrolled as "informant" and active participants in the exploration of his/her behaviors.

Naturalness and the Problem of External Validity

External validity concerns the extent to which the findings of a study are generalizable beyond the specific domains of that study. The relevant questions are directed at generalizability across subjects, experimental conditions, time, settings, and behaviors. Intuitively, external validity is limited to the extent that the set of conditions under which the results were obtained differ from the set of conditions to which one wants to generalize. Moreover, if generalization to real-world conditions is desired, then the closer the experimental conditions resemble real-world conditions, the greater is the potential external validity. Focusing on these assertions, the artificiality and concomitant generalizability limitations of personality-oriented research with correctional populations are viewed as a central methodological problem.

Tunnel (1977) has recently conceptualized three independent dimensions of naturalness—setting, behavior, and treatment—that he argues should be incorporated into research designs to increase generalizability of experimental results. Tunnel defined these dimensions as follows: (a) natural setting = “a context outside the lab to which a person is naturally exposed” (p. 427); (b) natural behavior = “one that is not established or maintained for the sole or primary purpose of conducting research; the behavior is part of the person’s existing repertoire” (p. 426); and (c) natural treatment = “a natural, discrete event, temporally bounded, that would have occurred without the researcher’s presence” (p. 427).

Even though research with prisoners has quite often been conducted in natural settings, Tunnel’s (1977) natural behavior and natural treatment dimensions have rarely been used in research designs. For example, of the 28 personality discrimination studies listed in Table 1, only 3 used natural behaviors as dependent measures, and none used natural treatments as independent measures. Some of the most frequently used experimental tasks include answering hypothetical questions about delaying gratification, estimating when or if future events will occur, and responding

to true-false or agree-disagree personality items. Clearly, in these tasks, neither behaviors nor treatment bear much resemblance to real-life conditions. As Mischel (1977) so aptly put it:

In the conditions of real life, the psychological “stimuli” that people encounter are neither questionnaire items, nor experimental instructions, nor inanimate events, but involve people and reciprocal relationships. (p. 248)

As a consequence of ignoring these naturalistic dimensions, the results of correctional population research are severely limited in their generalizability across behaviors and to real-world situations. Again, Mischel’s (1977) comment accurately reflects our position:

The future of personality measurement will be brighter if we can move beyond our favorite pencil-and-paper and laboratory measures to include direct observation as well as unobtrusive nonreactive measures to study lives where they are really lived and not merely where the researcher finds it convenient to look at them. (p. 248)

An article recently reviewed for *JCCP* illustrates the artificiality problem and the concomitant difficulty of generalizing across experimental conditions. The goal of the research was to elucidate components of modeling techniques and juvenile offender types potentially important in the design of treatment programs. As independent variables, two offender types (immature inadequate and socialized subcultural) and two model status dimensions (peer and staff) were studied, with imitation of self-reward criteria on a pursuit rotor task as the dependent variable. Consistent with the predictions, interaction effects were obtained in which offender types were found to be differentially susceptible to status of model influences. The authors concluded that “the results of this present study have confirmed the efficacy of the application of modeling techniques to the treatment of juvenile offenders.” Although this study was relatively free from internal validity problems, the authors’ conclusion seemed unjustifiable from the standpoint of external validity. Given that a Model Status \times Offender Type interaction was demonstrated within the microculture of this research, the potential importance of these influences to the treatment process bring into focus questions of increased

complexity. For example, the nature of Model Status \times Offender Type interactions may be quite dependent on broader social contextual influence in the natural environments of juvenile offenders. Certainly a particular setting (e.g., a correctional facility) may have considerable influence on what type of offender will imitate what type of model. Even though it is recognized that this criticism is rather pervasive and certainly not unique to this specific study, the authors provided no indication of an awareness of the complex nature of these issues.

A second external validity problem highlighted by this study centered on the choice of dependent measures. The imitation of self-reward criteria on a pursuit rotor task is quite distant from treatment goals relevant to juvenile offenders. Yet the investigators, as is often the case, suggested the extension of their method to treatment on the basis of the experimental results. Even with a more face-valid dependent measure (e.g., a treatment-relevant activity such as an increase in positive peer interactions), demonstrated generalizability across situations should be a goal.

Psychological research with correctional populations has unfortunately concentrated on artificial rather than naturalistic paradigms and measures. As a result, external validity issues have been largely neglected. As with most psychological investigations, the focus has been on internal validity issues. This underscores the internal versus external validity tradeoff in which more restrictive controls to deal with internal validity problems have reduced generalizability and increased external validity problems. Since the ultimate aim of research with correctional populations is generalizability to natural behaviors and situations, more attention should be paid to external validity issues even if it means initial loss of confidence in experimentally derived relationships. To be able to record accurately all responses in a pursuit rotor task provides us with little or no information regarding naturally occurring behavior. A shift in focus to more natural behaviors and treatments should also function as a catalyst for the development of improved methods for dealing with internal validity problems under conditions of naturalness.

Recommendation

In an effort to increase external validity, psychologists should incorporate both natural treatments and natural behaviors, as well as natural settings, into research designs. The use of direct observation in natural settings and the concomitant search for and use of nonreactive, unobtrusive measures should be encouraged.

Recidivism: A Measure in Need of Refinement

Although there has been much written in the criminology/sociology literature (e.g., Adams, 1974; Empey & Erickson, 1972; Kirby, 1954; Lerman, 1968) regarding problems inherent in measures of recidivism (an index of renewed criminal behavior following contact with the criminal justice system), we were unable to locate any discussion of the limitations of this measure in a psychological journal. Yet, recidivism, defined in its most usual form as *reincarceration following release from a correctional institution*, was used as both an independent variable in several of the personality discrimination studies and a dependent measure in the adjustment prediction and treatment evaluation studies. In the former, subjects were classified as recidivists or nonrecidivists, after which their performance on various dependent measures of personality (obtained during their previous incarceration) was assessed in an effort to determine which measures discriminated between the groups. In the latter types of studies, a host of demographic and personality variables were used as predictors of recidivism. There was, however, no indication that the investigators had any misgivings about or were concerned with the methodological problems of the recidivism measure itself. The following is a brief discussion of several issues that should concern psychological researchers who use this measure.

A major problem for recidivism measures in both types of studies is the criminal behavior/system discretion confound. In essence, recidivism, however defined, involves discretionary judgments of criminal justice system personnel that cannot be extricated from the

criminal behavior dimension that this measure is ideally designed to tap. From the decision to arrest by the police to the decision to reincarcerate by the judiciary or parole official, several discretionary judgments are made. At the point of reincarceration, the summative impact of these judgments maximizes the likelihood that differential biases have operated in determining who becomes a recidivist. Some offenders (e.g., nonwhite lower-class urban youth) may be arrested and returned to a correctional facility for relatively minor infractions, whereas others (e.g., white middle-class suburban youth) may commit the same or more serious crimes and remain out on parole. Moreover, although recidivism rates are sensitive to policy shifts at all levels, the sensitivity increases the further one becomes entangled in the criminal justice system process, with the point of reincarceration allowing for the maximum policy and administrative impact. Sellin (1962), a well-known criminologist, has commented that

the difficulty with statistics drawn from later stages in the administrative process is that they may show changes or fluctuations which are not due to changes in criminality but to variations in the policies or the efficiencies of administrative agencies. (p. 64)

One of the best examples of problems engendered by the criminal behavior/system discretion confound comes from the California Youth Authority's investigation of the community treatment program (Warren, Neto, Palmer, & Turner, 1966). First-commitment youths were randomly assigned either to experimental services in their own communities or to a control group situation that involved residence in an institution away from home. The results indicated that in general, a significantly higher percentage of control than experimental youths had engaged in parole violations (e.g., parole was officially revoked, recommitment occurred, or unfavorable discharge from the Youth Authority was given).

Although these results may have supported the contention that the community-based services were superior, Lerman's (1968) reanalysis of the data suggested that the parole violation rates were quite misleading about the behavior of the two groups. For example, despite the higher parole violation rates for the control youths, the experimentals engaged

in more known delinquent offenses per boy than did the controls. Moreover, a consideration of seriousness levels revealed that the experimentals exhibited higher delinquent behavior rates for low- and medium-seriousness offenses, with both groups about equal with regard to high-seriousness offenses. This finding, however, was confounded by evidence that the parole officers of the experimental boys were much more likely to know about their boys' offenses than the parole officers of the controls. Thus, the delinquent behavioral output may have been the same for the two groups, but the rate of being noticed was significantly higher for the experimental boys (Lerman, 1968) and the rate of reincarceration was higher for the controls. The experimental youths thus had lower parole violation rates but higher rates of recorded delinquent offenses.

This apparent, albeit spurious, contradiction can be explained in terms of differential reactions of the experimental and control organizations to recorded offenses. Although the parole officers of the experimental boys were much more likely to notice offenses, they were also much *less* likely to handle these cases via the parole violation method except with high-seriousness offenses. The lower parole violation rates obtained by the experimental group may therefore be due to their parole officers' disinclination to handle medium- and low-seriousness offenses by parole violation. Thus, the measure of recidivism may actually be telling us more about the behavior of the parole officers than that of the boys.

As elucidated in this example, the policies (and concomitant process) for both detection and handling of offenses may differ across specific organizations and with regard to particular offenders. This difficulty combined with the necessarily judgmental nature of these tasks may engender serious problems for studies that depend on recidivism measures.

The criminal behavior/system discretion confound can also be examined from the perspective of Type I and Type II errors and their varying probabilities throughout the criminal justice system process. Although it is unfortunate for researchers concerned with the reliability of recidivism data, individuals are understandably reluctant to admit when

they commit a crime. Therefore, recidivism is known only when a person is arrested, convicted, sentenced, and/or institutionalized. Throughout this process of formal contact with the criminal justice system, both kinds of classification errors are possible. A Type I error, or false positive, occurs when a person is classified as a recidivist when he/she did not, in fact, engage in criminal behavior. A Type II error, or false negative, occurs when a person is classified as a nonrecidivist when he/she did, in fact, commit another offense following release from prison. The relative probability of each error type will vary with the stage of the criminal justice system process. At the earliest point in the process—the point of rearrest by the police—there is a much greater probability of Type I errors than at later stages. For example, when the process reaches reincarceration, Type I errors are minimized by the abundance of legal safeguards along the way (unless the person is still on parole and therefore can often be reincarcerated without any judicial review). These safeguards are, of course, designed to prevent a person from erroneously being sent back to prison. Although the probability of Type I errors decreases as one moves further into the administrative process, the probability of Type II errors correspondingly increases. Thus, the Type II error presents the major problem for studies that define recidivism as reincarceration.

Type II errors present at least two kinds of relevant problems. In each case the difficulty lies with the determination of the nonrecidivist group. First, due to the impossibility of detecting every offender and the inconsistencies of crime detection systems, many persons classified as nonrecidivists may, in fact, have committed a crime. In the evaluation of the community treatment program, for example, the fact that the experimental boys had more recorded offenses than the controls may have been due to a greater likelihood of being noticed by their parole officers rather than to any real difference in delinquent behavioral output. Detection system differences rather than behavioral differences may thus play a role in the recidivist-nonrecidivist classification process. Second, with recidivism defined as reincarceration, all parole viola-

tions and other forms of criminal justice system contact short of reincarceration may still result in a person being classified as a nonrecidivist. Fortunately, most of the studies examined defined a nonrecidivist as a person having *no contact* with the criminal justice system. A problem still remains, however. Although the no-contact and reincarceration criteria for nonrecidivist and recidivist may assure maximum differences between groups (and maximum differences in criminal justice system reentanglement), all gradations of law-violative behavior and criminal justice system contact in between these extremes is ignored. Some investigators (e.g., Roberts, Erikson, Riddle, & Bacon, 1974) improved their methodology by including all of these gradations into a third, *some contact*, group. Finer differentiations (e.g., by frequency and type of offense) with more than three groups could yield still more accurate and complete information.

Two additional problems, both related to institutional bookkeeping, warrant attention. First, individuals are often counted as nonrecidivists who end up in other institutional programs (e.g., mental hospitals), die, or move out of state where they may engage in criminal behavior and even be reincarcerated. The point is that these individuals, because they have not been reincarcerated in a correctional facility in the same state, are often classified as nonrecidivist. Clearly this designation is erroneous. Second, although most investigators suggested that they used no contact with the criminal justice system as a criterion for nonrecidivism classification, in no study was there any indication of how this information was gathered. The problem is that most investigators usually consult only one source to gain this information under the assumption that there is a coordinated record keeping system between all aspects of the criminal justice apparatus within a given state. This is a faulty assumption. Not only does such a coordinated system usually not exist, but also depending on which source of information one uses (e.g., police records, correctional department files, parole officers reports, court records), one is likely to find different numbers and types of recidivism information on any given individual. Thus, dif-

ferent conclusions about the extent and nature of recidivism may vary depending on source of data. Although these problems may be more serious when recidivism is used as a measure of treatment program success, researchers who use this measure in other types of studies should be aware of the fact that the measure is not as straightforward and internally valid as it is generally assumed.

Recommendation

Recidivism should be operationally defined whenever it is used, and source of data should be provided. When used as a measure of post-institutional adjustment, recidivism should be supplemented by other measures of adjustment in the areas of education, employment, and family life. When possible, recidivism measures should be based on multiple gradations of such factors as extent of contact with the criminal justice system, frequency of law-violative behavior, and type and seriousness of offense.

Ethics of Research with Correctional Populations

This is an area that has attracted much professional attention recently but one that some investigators may not wish to consider as a methodological issue. We, however, consider it to be so integrally related that a brief discussion is warranted within the present context.

The National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research (1977) has recently provided recommendations regarding prison research that have considerable merit. The Commission recognized three broad categories:

1. Research conducted with the goal of improving institutional and program effectiveness, which includes psychological treatments having "the intent or reasonable probability of improving the health or well-being of the individual prisoner" (p. 3080);

2. Research inherently related to prisoners but not having a goal of benefiting them (e.g., investigations of prisoner personality) or "the possible causes, effects and processes of incarceration and studies of prisons as in-

stitutional structures or of prisoners as incarcerated persons" (p. 3080); and

3. Research using prisoners as subjects because they are available rather than because of their status as prisoners (e.g., psychopharmaceutical testing).

The Commission recommended that a human subjects review committee, composed of individuals of diverse racial and cultural backgrounds, including prisoners, prisoner advocates, clergy, community representatives, behavioral scientists, and medical personnel, should approve *all* research. This committee would consider the risks involved in the research, the provisions for obtaining informed consent, and safeguards to protect confidentiality and other concerns. Furthermore, parole boards could not take into account prisoners' participation in research. Research in the second category, that is, nonbeneficial but related to prisoners, must additionally "present minimal or no risk and no more than mere inconvenience to the subjects" (National Commission for the Protection, 1977, p. 3080). Finally, research on prisoners because of their availability could only be conducted if three additional requirements were met: (a) The research must fill an "important social and scientific need," (b) it must "satisfy conditions of equity," and (c) it must be distinguished by "a high degree of voluntariness" on the part of the prisoner.

These recommendations deserve close attention and adherence. As Brodsky (Note 2) recently stated in his report to the APA Task Force on Psychology and the Criminal Justice System:

In spite of the fact that there may be some observational and nonreactive investigations which would seem to have minimal impact—such as questionnaire studies—the potential for both misuse and harm to individuals exists in any study. In these as well as the more active manipulations of treatment conditions or offenders, it is important that consent forms, institutional reviews and other elements of protection for human subjects be included. (p. 57)

Recommendation

Some acknowledgment that ethical safeguards have been used (e.g., informed consent was obtained) should be included in the Method section of all research submitted for publication in APA journals.

Conclusion

In concluding this article it is perhaps worth stating that the issues discussed are neither exhaustive of the methodological problems confronting psychologists doing research with correctional populations nor are they complete discussions in and of themselves. Moreover, the emphasis on personality discrimination research may provide the reader with a distorted view of the type of research that is conducted with correctional populations. Although this may be true, it is certainly no distortion to suggest that this is the predominant type of research that is published in *JCCP* and *JAP*. Finally, we do not expect that all of the issues discussed can be satisfactorily resolved in any single research project. Nevertheless, we firmly believe that most of these issues have received little sustained attention in past psychological research with correctional populations and that this state of affairs should not continue.

Reference Notes

- West, S. G., & Gunn, S. G. *Some issues of ethics and societal psychology*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C., August-September 1976.
2. Brodsky, S. L. Ethical issues for psychologists in corrections. In J. Monahan (Chair), *Unpublished report of the Task Force on the Role of Psychology in the Criminal Justice System*. Washington, D.C.: Board of Social and Ethical Responsibility for Psychology, American Psychological Association, 1978.
3. Brody, P. R. *Manipulability of predisposition to choose immediate or delayed gratification*. Unpublished manuscript, Yale University, 1967.

References

- Adams, S. Measurement of effectiveness and efficiency in corrections. In D. Glaser (Ed.), *Handbook of criminology*. Chicago, Ill.: Rand McNally, 1974.
- Allport, G. W. *Personality: A psychological interpretation*. New York: Holt, 1937.
- Allport, G. W. The psychologists' frame of reference. *Psychological Bulletin*, 1940, 37, 1-28.
- Allport, G. W. The ego in contemporary psychology. *Psychological Review*, 1943, 50, 451-478.
- Argyle, M., & Little, B. R. Do personality traits apply to social behavior? *Journal for the Theory of Social Behavior*, 1972, 2, 1-35.
- Armentrout, J. A., & Rouzer, D. L. Utility of the Mini-Mult with delinquents. *Journal of Consulting and Clinical Psychology*, 1970, 34, 450.
- Barber, T. X., & Silver, M. J. Fact, fiction, and the experimenter bias effect: Pitfalls in data analysis and interpretation. *Psychological Bulletin Monograph*, 1968, 70(6, Pt. 2).
- Baron, R. M., Cowan, G., Ganz, R. L., & McDonald, M. Interaction of locus of control and type of performance feedback: Considerations of external validity. *Journal of Personality and Social Psychology*, 1974, 30, 285-292.
- Barratt, E. Anxiety and impulsiveness related to psychomotor efficiency. *Perceptual and Motor Skills*, 1959, 9, 191-198.
- Bem, D., & Allen, A. On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, 1974, 81, 506-520.
- Bixenstine, V. E., & Buterbaugh, R. L. Integrative behavior in adolescent boys as a function of delinquency and race. *Journal of Consulting and Clinical Psychology*, 1967, 31, 471-476.
- Blackburn, R. Sensation seeking, impulsivity, and psychopathic personality. *Journal of Consulting and Clinical Psychology*, 1969, 33, 571-574.
- Borkovec, T. D. Autonomic reactivity to sensory stimulation in psychopathic, neurotic, and normal juvenile delinquents. *Journal of Consulting and Clinical Psychology*, 1970, 35, 217-222.
- Bowers, K. S. Situationism in psychology: An analysis and critique. *Psychological Review*, 1973, 80, 307-337.
- Brodsky, S. L. *Psychologists in the criminal justice system*. Urbana: University of Illinois Press, 1973.
- Campbell, D. T. Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist*, 1960, 15, 546-553.
- Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Rand McNally: Chicago, 1963.
- Caplan, N., & Nelson, S. On being useful: The nature and consequences of psychological research on social problems. *American Psychologist*, 1973, 28, 199-211.
- Carlson, K. A. Classes of adult offenders: A multivariate approach. *Journal of Abnormal Psychology*, 1972, 79, 84-93.
- Cohen, M., Seghorn, & Calmas, W. Sociometric study of the sex offender. *Journal of Abnormal Psychology*, 1969, 74, 249-255.
- Cowden, J. E., & Pacht, A. R. Predicting institutional and postrelease adjustment of delinquent boys. *Journal of Consulting Psychology*, 1967, 31, 377-381.
- Cronbach, L. J. Beyond the two disciplines of scientific psychology. *American Psychologist*, 1975, 30, 116-127.

- Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, 52, 281-302.
- Cronbach, L. J., & Snow, R. E. *Aptitudes and instructional methods*. New York: Irvington, 1975.
- Davids, A., & Falkof, B. B. Juvenile delinquents then and now: Comparison of findings from 1959 and 1974. *Journal of Abnormal Psychology*, 1975, 84, 161-164.
- Davidson, W. S., & Seidman, E. Studies of behavior modification and juvenile delinquency: A review, methodological critique, and social perspective. *Psychological Bulletin*, 1974, 81, 998-1011.
- Emery, R. E., & Marholin, D. An applied behavior analysis of delinquency: The irrelevancy of relevant behavior. *American Psychologist*, 1977, 32, 860-873.
- Empey, L. T., & Erikson, M. L. *The Provo experiment: Evaluating community control and delinquency*. Lexington, Mass.: Lexington Books, 1972.
- Endler, N. S. The person versus the situation—A pseudoissue? A response to Alker. *Journal of Personality*, 1973, 41, 287-303.
- Endler, N. S. The case for person-situation interactions. *Canadian Psychological Review*, 1975, 16, 12-21. (a)
- Endler, N. S. A person-situation interaction model for anxiety. In C. D. Spielberger & J. G. Sarason (Eds.), *Stress and anxiety*. (Vol. 1). Washington, D.C.: Hemisphere, 1975. (b)
- Endler, N. S., & Hunt, J. McV. Sources of behavioral variance as measured by the S-R Inventory of Anxiousness. *Psychological Bulletin*, 1966, 65, 336-346.
- Endler, N. S., & Magnusson, D. Toward an interactional psychology of personality. *Psychological Bulletin*, 1976, 83, 956-974.
- Erikson, R. V., & Roberts, A. H. Some ego functions associated with delay of gratification in male delinquents. *Journal of Consulting and Clinical Psychology*, 1971, 36, 378-382.
- Ganzer, V. J., & Sarason, I. R. Variables associated with recidivism among juvenile delinquents. *Journal of Consulting and Clinical Psychology*, 1973, 40, 1-5.
- Glaser, D. Remedies for the key deficiency in criminal justice evaluation research. *Journal of Research in Crime and Delinquency*, 1974, 11, 144-154.
- Goffman, E. *Asylums*. New York: Doubleday Books, 1961.
- Grisso, J. T. Characteristics of institutionalized delinquents in conflict about their release. *Journal of Consulting and Clinical Psychology*, 1975, 43, 267.
- Gunn, J., & Gristwood, J. Use of the Buss-Durkee Hostility Inventory among British prisoners. *Journal of Consulting and Clinical Psychology*, 1975, 43, 590.
- Hare, R. D. Psychopathy, automatic functioning, and the orienting response. *Journal of Abnormal Psychology Monograph Supplement*, 1968, 73(3, Pt. 2).
- Hartshorne, H., & May, M. A. *Studies in the nature of character: Studies in deceit* (Vol. 1). New York: Macmillan, 1928.
- Heller, K., & Monahan, J. *Psychology and community change*. Homewood, Ill.: Dorsey Press, 1977.
- Hetherington, E. M., Stouwie, R. J., & Ridberg, E. H. Patterns of family interaction and child-rearing attitudes related to three dimensions of juvenile delinquency. *Journal of Abnormal Psychology*, 1971, 78, 160-176.
- Hirschfield, P. Response set in impulsive children. *Journal of Genetic Psychology*, 1965, 107, 117-126.
- Hudson, L. *The cult of the fact*. New York: Harper & Row, 1972.
- Jesness, C. F. Comparative effectiveness of behavior modification and transactional analysis. *Journal of Consulting and Clinical Psychology*, 1975, 43, 758-779.
- Jones, E. E., & Nisbet, R. E. *The actor and observer: Divergent perceptions of the causes of behavior*. New York: General Learning Press, 1971.
- Jourard, S. M. *Disclosing man to himself*. Princeton, N.J.: Van Nostrand, 1968.
- Jurkovic, G. J., & Prentice, N. M. Relation of moral and cognitive development to dimensions of juvenile delinquency. *Journal of Abnormal Psychology*, 1977, 86, 414-420.
- Kercher, G. A., & Walker, C. E. Reactions of convicted rapists to sexually explicit stimuli. *Journal of Abnormal Psychology*, 1973, 81, 46-51.
- Kirby, B. C. Measuring effects of treatment of criminals and delinquents. *Sociology and Social Research*, 1954, 38, 368-375.
- Kulik, J. A., Stein, K. B., & Sarbin, T. R. Dimensions and patterns of adolescent antisocial behavior. *Journal of Consulting and Clinical Psychology*, 1968, 32, 375-382. (a)
- Kulik, J. A., Stein, K. B., & Sarbin, T. R. Disclosure of delinquent behavior under conditions of anonymity and non-anonymity. *Journal of Consulting and Clinical Psychology*, 1968, 32, 506-509. (b)
- Landau, S. F. Delinquency, institutionalization, and time orientation. *Journal of Consulting and Clinical Psychology*, 1976, 44, 745-759.
- Lerman, P. Evaluative studies of institutions for delinquents: Implications for research and social policy. *Social Work*, 1968, 13, 55-64.
- Levine, M. Scientific method and the adversary model: Some preliminary thoughts. *American Psychologist*, 1974, 29, 661-667.
- Magnusson, D., & Ekehammar, B. Perceptions of and reactions to stressful situations. *Journal of Personality and Social Psychology*, 1975, 31, 1147-1154.
- McArthur, A. V. *Coming out cold*. Lexington, Mass.: Heath, 1974.
- Magnusson, D., Heffler, B., & Nyman, B. The generality of behavioral data: Replication of an experiment on generalization from observation on one occasion. *Multivariate Behavioral Research*, 1968, 3, 415-422.
- Mariotti, M. J., & Paul, G. L. Persons versus situations in the real-life functioning of chronically institutionalized mental patients. *Journal of Abnormal Psychology*, 1975, 84, 483-493.

- McCandless, B. R., Persons, W. S., III, & Roberts, A. Perceived opportunity, delinquency, race, and body build among delinquent youth. *Journal of Consulting and Clinical Psychology*, 1972, 38, 281-287.
- McGuigan, F. J. The experimenter: A neglected stimulus object. *Psychological Bulletin*, 1963, 60, 421-428.
- McGuire, J. S., & Megargee, E. I. Personality correlates of marijuana use among youthful offenders. *Journal of Consulting and Clinical Psychology*, 1974, 42, 124-133.
- Megargee, E. The prediction of violence with psychological tests. In C. Spielberger (Ed.), *Current topics in clinical and community psychology*. New York: Academic Press, 1970.
- Megargee, E. I., Cook, P. E., & Mendelsohn, G. A. Development and validation of an MMPI scale of assaultiveness in overcontrolled individuals. *Journal of Abnormal Psychology*, 1967, 72, 519-528.
- Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, 30, 955-966.
- Mischel, W. Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, 1973, 80, 252-283.
- Mischel, W. On the future of personality measurement. *American Psychologist*, 1977, 32, 246-254.
- Monahan, J. The prevention of violence. In J. Monahan (Ed.), *Community mental health and the criminal justice system*. New York: Pergamon Press, 1976.
- Moos, R. H. Situational analysis of a therapeutic community milieu. *Journal of Abnormal Psychology*, 1968, 73, 49-61.
- Moos, R. H. Sources of variance in response to questionnaires and in behavior. *Journal of Abnormal Psychology*, 1969, 74, 405-412.
- Moos, R. H. *Evaluating correctional and community settings*. New York: Wiley, 1975.
- Mosher, D. L., Mortimer, P. L., & Grebel, M. Verbal aggressive behavior in delinquent boys. *Journal of Abnormal Psychology*, 1968, 73, 454-460.
- Moss, C. S., Hosford, R. E., Anderson, W. R., & Petracca, M. Personality variables of blacks participating in a prison riot. *Journal of Consulting and Clinical Psychology*, 1977, 45, 505-512.
- National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research. Department of Health, Education and Welfare. Protection of human subjects: Research involving prisoners—Report and recommendations. *Federal Register*, 1977, 42, 3075-3091.
- Oliver, W. A., & Mosher, D. L. Psychopathology and guilt in heterosexual and subgroups of homosexual reformatory inmates. *Journal of Abnormal Psychology*, 1968, 73, 323-329.
- Orris, J. B. Visual monitoring performance in three subgroups of male delinquents. *Journal of Abnormal Psychology*, 1969, 74, 227-229.
- Persons, R. W., & Marks, P. A. Self-disclosure with recipients: Optimum interviewer-interviewee matching. *Journal of Abnormal Psychology*, 1970, 76, 387-391.
- Rausch, H. L., Farbman, I., & Llewellyn, L. G. Person, setting, and change in social interaction: A normal control study. *Human Relations*, 1960, 13, 305-333.
- Redfering, D. L. Durability of effects of group counseling with institutionalized delinquent females. *Journal of Abnormal Psychology*, 1973, 82, 85-86.
- Roberts, A. H., & Erikson, R. V. Delays of gratification, Porteus Maze test performance, and behavioral adjustment in a delinquent group. *Journal of Abnormal Psychology*, 1968, 73, 449-453.
- Roberts, A. H., & Erikson, R. V. Measuring impulse control in institutionalized delinquents using Rorschach content and thought process scales. *Journal of Abnormal Psychology*, 1969, 74, 632-634.
- Roberts, A. H., Erikson, R. V., Riddle, M., & Bacon, J. G. Demographic variables, base rates, and personality characteristics associated with recidivism in male delinquents. *Journal of Consulting and Clinical Psychology*, 1974, 42, 833-841.
- Rosenthal, R. Experimenter attributes as determinants of subject responses. *Journal of Projective Techniques and Personality Assessment*, 1963, 27, 324-331.
- Rosenthal, R. Experimenter expectancy and the reassuring nature of the null hypothesis decision procedure. *Psychological Bulletin*, 1968, 70, 30-47.
- Rotenberg, M., & Sarbin, T. R. Impact of differentially significant others on role involvement: An experiment with prison social types. *Journal of Abnormal Psychology*, 1971, 77, 97-107.
- Ruma, E. H., & Mosher, D. L. Relationship between moral judgment and guilt in delinquent boys. *Journal of Abnormal Psychology*, 1967, 72, 122-127.
- Rumenik, D. K., Capasso, D. R., & Hendrick, C. Experimenter sex effects in behavioral research. *Psychological Bulletin*, 1977, 84, 852-877.
- Rushton, J. P. Socialization and the altruistic behavior of children. *Psychological Bulletin*, 1976, 83, 898-913.
- Sarason, S. B. *Psychology To the Finland Station in The Heavenly City of the Eighteenth Century Philosophers*. *American Psychologist*, 1975, 30, 1072-1080.
- Sarri, R. C., & Selo, E. Evaluation process and outcome in juvenile corrections: Musings on a grim tale. In P. O. Davidson, F. W. Clark & L. A. Hamerlynck (Eds.), *Evaluation of behavioral programs in community, residential, and school setting*. Champaign, Ill.: Research Press, 1974.
- Saunders, J. J., Reppucci, N. D., & Sarata, B. P. V. An examination of impulsivity as a trait characterizing delinquent youth. *American Journal of Orthopsychiatry*, 1973, 43, 789-795.
- Schlichter, K. J., & Ratliff, R. G. Discrimination learning in juvenile delinquents. *Journal of Abnormal Psychology*, 1971, 77, 46-48.
- Schmid, A. L. Susceptibility to social influence and retention of opinion change in two types of delinquents. *Journal of Abnormal Psychology*, 1970, 76, 123-129.
- Schwitzgebel, R. L. Preliminary socialization for psychotherapy of behavior-disordered adolescents.

- Journal of Consulting and Clinical Psychology*, 1969, 33, 71-77.
- Sellin, T. The significance of records of crime. In M. E. Wolfgang, L. Savitz, & N. Johnson (Eds.), *The sociology of crime and delinquency*. New York: Wiley, 1962.
- Silverman, I. The experimenter: A (still) neglected stimulus object. *Canadian Psychologist*, 1974, 15, 258-270.
- Silverman, I., Shulman, A. D., & Wiesensthal, D. L. The experimenter as a source of variance in psychological research: Modeling and sex effects. *Journal of Personality and Social Psychology*, 1972, 21, 219-227.
- Skrzypek, G. J. Effect of perceptual isolation and arousal on anxiety, complexity preference, and novelty preference in psychopathic and neurotic delinquents. *Journal of Abnormal Psychology*, 1969, 74, 321-329.
- Sowles, R. C., & Gil, J. H. Institutional and community adjustment of delinquents following counseling. *Journal of Consulting and Clinical Psychology*, 1970, 34, 398-402.
- Stein, K. B., Sarbin, T. R., & Kulik, J. A. Future time perspective: Its relation to the socialization process and the delinquent role. *Journal of Consulting and Clinical Psychology*, 1968, 32, 257-264.
- Stewart, D. J. Effects of social reinforcement on dependency and aggressive responses of psychopathic, neurotic, and subcultural delinquents. *Journal of Abnormal Psychology*, 1972, 79, 76-83.
- Stewart, D. J., & Resnick, J. H. Verbal conditioning and dependency behavior in delinquents. *Journal of Abnormal Psychology*, 1970, 76, 375-377.
- Street, D., Vinter, R. D., & Perrow, C. *Organization for treatment: A comparative study of institutions for delinquents*. New York: Free Press, 1966.
- Sutker, P. B. Personality differences and sociopathy in heroin addict and nonaddict prisoners. *Journal of Abnormal Psychology*, 1971, 78, 247-251.
- Truax, C. B., Wargo, D. G., & Volksdorf, N. R. Antecedents to outcome in group counseling with institutionalized juvenile delinquents. *Journal of Abnormal Psychology*, 1970, 76, 235-242.
- Tunnel, G. B. Three dimensions of naturalness: An expanded definition of field research. *Psychological Bulletin*, 1977, 84, 426-437.
- Warren, M. Q., Neto, V. V., Palmer, T. B., & Turner, J. K. *Community treatment project: An evaluation of community treatment for delinquents* (CTP Research Report No. 7). Sacramento: California Youth Authority, Division of Research, 1966.
- Wenk, E., Robison, T., & Smith, G. Can violence be predicted? *Crime and Delinquency*, 1972, 18, 393-402.
- Wilson, D. W., & Donnerstein, E. Legal and ethical aspects of nonreactive social psychological research. *American Psychologist*, 1976, 31, 765-773.
- Zimbardo, P. G. The mind is a formidable jailer: A Pirandellian prison. *New York Times Magazine*, April 8, 1973, p. 38.

Received January 30, 1978 ■

Methodological Errors in Marital and Child Treatment Research

K. Daniel O'Leary and Hillary Turkewitz
State University of New York at Stony Brook

Common methodological errors in child and marital treatment research are discussed, and suggestions are made to help investigators avoid such errors. The following areas are covered: selection of subjects and therapists, scope and source of dependent measures, treatment specification, experimental design, and data analysis and interpretation. Some of the most salient errors include (a) unsubstantiated diagnoses or client labels; (b) very few therapists per treatment condition; (c) restricted outcome criteria and the lack of reliable, valid dependent measures; (d) failure to provide treatment manuals and to check empirically whether the treatments were actually implemented; and (e) experimental designs that fail to address issues such as maturation, expectation, nonspecific relationship factors, and practical significance.

Our purpose is to discuss methodological problems that are common to most child and marital treatment research. Child and marital therapy are addressed as separate activities, although we recognize that childhood disorders are often influenced by marital discord and vice versa. That is, even though the problems of children and marriage are often not independent, we are reviewing the difficulties in research when either the child or the marital interaction is the focus of treatment. Such specific programs have proven successful in treating marital (Gurman & Kniskern, 1978) and childhood problems (K. D. O'Leary & Wilson, 1975). We recognize that individuals with other treatment orientations might view child and marital problems only within a family context and would treat the whole family—children, parents, and sometimes grandparents—simultaneously. Family therapy research will not be specifically addressed here; readers interested in this methodology should consult Wells, Dilkes, and Burckhardt (1976).

Marital discord has highly diverse etiologies, and distressed couples present a wide variety of problems including role conflicts, jealousy, individual pathology, physical abuse, and sexual dissatisfaction. Despite the varied presenting problems, most marital researchers and therapists have the commonly accepted goals of enhancing communication (D. H. Olson, 1970) and increasing satisfaction. The outcome research involving these general goals will be the focus of this article. Those interested in the research methodology of treatment for specific sexual dysfunctions should consult LoPiccolo (1978).

Two of the most commonly diagnosed childhood problems are conduct disorders (unsocialized aggressive reaction) and hyperkinesis (Cerreto & Tuma, 1977), and we will illustrate methodological errors in child treatment research from these areas. Both problems are the result of complex psychological, social, and biological factors, and the goals of treatment are multiple, including increases in academic productivity, frustration tolerance, attention span, positive parent-child interactions, and self-esteem. Consequently, the childhood problems covered represent areas of concern to most child treatment researchers. For those interested in other specific childhood problems, such as autism, enuresis,

We are especially appreciative of the editorial comments of S. G. O'Leary, D. Gelfand, and A. Gurman. Requests for reprints should be sent to K. Daniel O'Leary, Department of Psychology, State University of New York, Stony Brook, New York 11794.

adolescent suicide, and delinquency, *Annual Progress in Child Psychiatry and Child Development* (Chess & Thomas, 1977) and *Handbook of Treatment of Mental Disorders in Childhood and Adolescence* (Wolman, Egan, & Ross, 1978) are recommended as sources of discussions on substantive and methodological problems.

We will address common methodological problems by noting frequent errors of omission and commission and by making suggestions that can help investigators avoid such errors. The methodological errors will be discussed in the order of their occurrence as one reads a manuscript (American Psychological Association, 1974). More specifically, we will discuss (a) subjects, (b) therapists, (c) dependent measures, (d) treatment specification, (e) experimental design, and (f) data analysis and interpretation of results. The suggestions are offered as guidelines, but it is recognized that in outcome research with child and marital populations, meeting all of the suggested recommendations may be impossible. The extent to which they are met, however, will clearly increase both the internal and external validity of the research.

Subjects

Small Sample Size

Controlled research in child, and particularly marital, therapy is plagued by the problem of small subject samples. Enough subjects should be included in each experimental group to enable an investigator to detect outcome differences of practical and/or theoretical import. If one desires a guaranteed degree of power, the necessary sample size can be determined empirically by using prior estimates of treatment effectiveness (Winer, 1971). In any case, as Hartmann, Roper, and Gelfand (1977) noted, one must consider subject attrition, since the sample size of concern is the terminal sample, that is, the number of individuals who complete *all* phases of a study from pretreatment to follow-up. Obtaining significant results with a small sample may be a compelling testimony to treatment effects, since small *ns* decrease power or the ability to detect treatment differences. How-

ever, the smaller the sample, the greater the probability of sampling error and artifacts that would limit the generalizability of the findings.

YAVIS Samples

Williams (1956) coined the acronym YAVIS (young, attractive, verbal, intelligent, successful) in his discussion of the 'typical psychotherapy patient. This problem of restricted patient populations is particularly salient in the controlled research on marital therapy. Almost all of the published studies contain young, well-educated, middle-class couples. Because age and/or number of years married can affect the outcome of different approaches to marital therapy (Turkewitz & O'Leary, Note 1), a sample should be selected that includes a wide age range or one's conclusions must be restricted to the particular age group involved in the study. Further research is necessary to evaluate marital treatment programs for both older clients and clients from lower socioeconomic backgrounds.

Unspecified Subject Characteristics

A very common error of omission in the presentation of research on child and marital therapy is inadequate sample description. In addition to age, educational level, occupation, socioeconomic status, and information regarding previous therapy, authors should specify the fee paid, if any, for treatment in the research program. In marital therapy outcome studies, the mean and range of years married, the number of children, and the number of previous marriages should also be reported. In child research, it is advisable to provide information on the parents' marital status and the involvement of mothers and/or fathers in treatment. Depending on the particular child problem under study, a thorough assessment and description of the sample may involve a measure of intellectual functioning, academic performance, or marital satisfaction of the parents. A complete description of subject characteristics is necessary so that other researchers can attempt replications of the research and so that practicing clinicians can judge if the treatment

program described is likely to have similar effects for the population with whom they are working.

Unsubstantiated Labels and Diagnoses

In addition to specifying demographic characteristics, researchers executing outcome studies generally describe the subjects as having a particular clinical problem, for example, marital distress, adjustment reaction of childhood, or hyperactivity. However, outcome research is often hampered by very general definitions of these labels. To allow for generalizations regarding the treatment and replications of studies, specific data substantiating the diagnoses must be provided. For example, in marital research, couples referred by divorce courts are probably more severely distressed than those referred by clergy. Degree of distress could also depend on geographical location and community sanctions regarding divorce. Obtaining a numerical rating of distress from a valid assessment instrument that has normative data provides a description of the sample that can be understood by other professionals and increases the reliability of the diagnosis of marital distress. The Locke-Wallace Marriage Relationship Inventory (Kimmel & Van der Veen, 1974; Locke & Wallace, 1959) is a useful questionnaire for these purposes, as discriminant validity (Locke & Wallace, 1959), construct validity (Weiss, Hops, & Patterson, 1973), and test-retest reliability (Kimmel & Van der Veen, 1974) have been demonstrated for various forms of this inventory.

Similarly, the need for specific, defined criteria is particularly salient in the research on hyperactivity. Children may be diagnosed as hyperactive by a pediatrician, clinical psychologist, psychiatrist, or school psychologist. Additionally, children may receive the label because they are taking Ritalin, have been referred by a teacher, or have parents who sought help in handling them at home. These varying sources of labels yield highly heterogeneous groups of children described as hyperactive. In this regard, the prevalence rates of hyperactivity range from 4% to 20% of elementary school samples (Ross & Ross, 1977). Even though teachers are probably

the best information source regarding a child's activity level, caution should be observed in basing a diagnosis of hyperactivity on a single teacher's referral. Teachers occasionally inflate their assessment of a child's problems because they want to obtain consultation from the researcher (S. O'Leary, Note 2). Further, mental health practitioners may not be able to diagnose hyperactivity reliably (Kenny et al., 1971). A recommended assessment measure for hyperactive children is the Conners Teacher Rating Scale (Conners, 1969). However, such a measure should be supplemented by parent ratings (Routh, Schroeder, & Otuama, 1974). The general principle is that labeling or diagnosing the subject population requires well-validated, reliable measures to substantiate the descriptive label.

Inadequate and Unspecified Selection Process

An investigator conducting outcome research should define the problem under study and include only those subjects who meet a priori criteria. The use of a priori criteria increases the likelihood that the sample obtained has clinically significant problems and will be comparable to those in other investigations. (This argument assumes that some other investigators also use commonly accepted criteria for clinical problems.)

We recognize that it is often difficult to obtain a sufficiently large sample if one uses stringent a priori criteria regarding severity of the problem. If this is not possible, a more general criterion such as referral by a teacher, parent, pediatrician, mental health clinic, or divorce court should be accompanied by a specific description of the referral process and descriptive measures like the ones discussed in the previous section. If more than one referral source is used, the percentage of subjects obtained from each different source should be noted, as different sources can produce clinically different samples, for example, teacher versus pediatrician and clinic versus court referrals.

In addition to a severity criterion, many researchers exclude clients who would not be appropriate for the treatment under study. Examples of these exclusionary criteria in

short-term marital therapy are evidence of severe psychopathology in either spouse (e.g., psychosis or chronic alcoholism), concurrent therapy, spouses living separately, and the refusal of either spouse to cooperate in the treatment program. In child treatment programs, exclusion has been based on neurological impairment, childhood psychosis, inability of both parents to attend therapy sessions, severe marital discord, or uncooperative teachers. Although these exclusionary criteria may be valid for certain treatments, they are often not specified. For example, researchers rarely state how they determined the presence of neurological involvement or severe psychopathology. As excluding subjects with these problems limits claims of generalizability of the results, it is important to describe the decision-making process and specify the percentage of potential subjects who were excluded.

Therapists

Small Therapist Samples

One of the most common methodological errors made in marital and child outcome research is that the number of therapists involved is too small to allow for appropriate generalizations regarding treatment. Therapist factors have been repeatedly shown to influence outcome in family therapy (Gurman & Kniskern, 1978), and it is very likely that these factors also have an impact on child and marital treatment. Given the importance of therapist variables, it is necessary to have as many therapists as practically feasible so that one can study whether one's treatment program can be successfully implemented by therapists of varying styles. Keeping in consideration the need for training and close supervision in studies of particular treatment programs, a minimum of 3-4 therapists per experimental condition is strongly advised. When therapist factors become a major focus of the research, the number of therapists should be even greater.

Bias Introduced by Motivational Factors

Teacher incentives. In child treatment research, motivational factors need to be con-

sidered when teachers are asked to implement therapeutic programs in their classrooms. If a teacher receives graduate credits, free time, additional classroom materials, or teaching assistants for conducting an intervention, these potential reinforcers can positively bias the consistency and quality of the teacher's efforts. We would not advise against using such incentives, but it should be noted that clinicians in nonresearch settings cannot offer the teacher incentives and consequently might not find the intervention to be as successful as it appeared in the research program.

Authors as therapists. A serious methodological problem regarding therapist selection involves the experimenter(s) as therapist(s). The emotional investment and enthusiasm of the author/therapist may increase placebo effects and spuriously inflate the success of the treatment under evaluation. These factors would hold even if there were two or three therapists, if all were authors.

Confounds Resulting from Therapist Characteristics or Theoretical Biases

The general issue of motivation and enthusiasm for a particular treatment approach becomes even more critical when two or more treatments are being compared. Many of the comparative studies in the marital and child treatment areas involve two or more therapists who implement both treatments under study. This methodology of crossing therapists with treatments has clear advantages and disadvantages. One advantage is that differential treatment outcomes can more readily be attributed to the treatments rather than to the skill of the therapists (since the therapists are the same for both treatment programs). One serious disadvantage, however, is the possibility of systematic bias if all the therapists involved have a theoretical orientation that favors one of the treatments. For example, if several behavior therapists were to execute a study comparing behavioral marital therapy with a systems approach and if these behavior therapists were to conduct both treatment programs, a bias would be introduced by greater enthusiasm for the behavioral intervention, greater knowledge of the specific behavioral procedures, and more

clinical experience in treating clients with behavioral methods.

Rather than crossing therapists with treatments, different therapists can execute each program under study. If this methodology is used, large numbers of therapists must be used to avoid confounding treatment effects and differential therapist skills. The difficulty in obtaining a sufficient number of therapists of differing orientations often led investigators to cross therapists with treatments. Because this method is frequently used, the following recommendations are made to reduce possible biases: Include therapists of varying theoretical orientations; present cogent rationales and theoretical explanations for all treatments under study; offer extensive training programs, including readings, role playing, and pilot cases; and provide close supervision. Even with these precautions, any therapist biases should be assessed and reported. For example, one can obtain the therapists' predictions regarding the effectiveness of the different approaches and their reports on how comfortable they were in conducting the treatments. The assessment and control for the biases introduced by therapist factors is important, because the small number of therapists often involved in treatment research makes it difficult, if not impossible, to study the influence of therapist characteristics statistically.

Inadequate Description of Therapists

Information should be included on the therapists' (and/or teachers') training, educational, and professional background; sex; age; previous experience with marital or child treatment; and incentives for participation (salaried vs. volunteer). In studies of marital and child therapy, the marital status of the therapists, and whether or not they have children may be of interest.

Dependent Measures

Multiple Outcome Criteria

A very serious error in child and marital treatment research is the lack of multiple outcome criteria (Johnson & Christensen,

1975). In both the child and marital areas, many investigators have relied exclusively on self-report measures. In contrast, some of the child treatment studies—especially evaluations of behavior therapy programs—have included only naturalistic observations. As self-report and observational measures both have strengths and weaknesses (K. D. O'Leary & Johnson, in press) and sometimes can lead to different conclusions (e.g., Harrell & Guernsey, 1976; Liberman, Levine, Wheeler, Sanders, & Wallace, 1976), neither data source should be considered sufficient. Each provides unique information that is necessary in any comprehensive treatment evaluation.

Some authors have argued for a tripartite assessment model including self-report, observational, and physiological measures (Lang, 1971). Such a model has been quite useful in evaluating fears and anxiety, but the model is not especially relevant to the evaluation of child and marital treatment. In child treatment, academic achievement tests, mechanical measures (such as stabilometer chairs to assess hyperactivity), and sociometric evaluations are useful adjuncts to self-report and observational measures. In marital treatment, adjunctive measures such as days absent from work, job productivity, and physical well-being could be considered.

Regardless of the type of assessment data obtained, a clinical interview is a critical and almost universal mode of assessment. However, the material collected in an interview, the meaning assigned to that material (e.g., a personality structure or a behavior reinforced by significant others), and the extent of standardization in the interview will depend on one's theoretical orientation. Under certain conditions a clinical interview can be reliable, and the interview allows for maximal flexibility in developing new hypotheses regarding a client's problem (K. D. O'Leary & Johnson, in press).

Comprehensive Assessment

In addition to the assessment modality that is used, there are two major issues to be considered when choosing dependent measures:

1. Scope of the assessment. For example in a child treatment study, possible factors

of interest include classroom academic and/or social behavior, self-esteem, a child's feelings toward his/her parents or teachers, and parent-child interactions at home. In marital therapy outcome research, one can consider assessment of individual functioning, sexual satisfaction, role conflicts, social life, expectations of and feelings toward one's spouse, and a consumer evaluation of the program.

2. Source of the data. Strupp and Hadley (1977) argued that psychotherapy outcomes should be evaluated from three different perspectives—(a) society or significant others (e.g., relatives, teachers, employers), (b) the client, and (c) the therapist. In particular, we emphasize the strong need for data from the child's perspective in evaluations of child therapy.

The choice of dependent measures and the data sources will be dictated by one's theoretical orientation and the particular questions of interest. It is beyond the scope of this article to discuss fully all of the factors that should be assessed in the study of various child or marital problems. (The reader is referred to Gurman and Kniskern, 1978, for a detailed discussion of this issue.) As self-report and observational measures are the most commonly used dependent measures in the evaluation of child and marital treatment, they will be discussed in detail.

Self-report

One of the most frequently cited problems of self-report measures is the influence of demand characteristics (e.g., cues that convey the experimental hypotheses and the implicit or explicit expectations to the clients). Demand characteristics can be minimized by keeping the specific criteria for inclusion in a treatment study unclear to the clients, by having the therapist absent when questionnaires are completed (either before or after therapy), and by taking great care not to pressure the clients to report success in order to help the therapist.

Social desirability, or the tendency to respond in a conventional, socially acceptable fashion, is also a problem that limits the utility of self-report data. For example, Edmonds (1967) found that a large percentage of

spouses answered "true" to *impossible* responses. (e.g., "There is never a moment I don't feel head over heels in love with my spouse.") It is moot whether these socially desirable responses reflect active distortions or honest tendencies of spouses to exaggerate positive qualities of their mates, and it has also been argued that a large percentage of socially desirable responses may not decrease the validity of a marital satisfaction test (Murstein & Beck, 1972). A full discussion of social desirability is not warranted here. However, when social desirability may be a critical factor (e.g., cases of child abuse), an investigator can consult the personality research literature for methods to control for it, such as transforming raw scores using a social desirability correction factor (e.g., Minnesota Multiphasic Personality Inventory) and developing special multiple-choice questionnaire formats.

A third problem of self-report measures that is often overlooked is that clients must have a certain language proficiency to complete questionnaires; and with few exceptions (e.g., Bienvenu, 1970), authors do not specify the degree of literacy required. When treating clients with limited educational backgrounds (a group much neglected in the marital research area), investigators should use a reading survey instrument to determine the language proficiency required to complete their questionnaire (e.g., Fry, 1968).

Finally, a common error is to use self-report measures that do not have established reliability and validity, which renders the degree of reported change very difficult, if not impossible, to interpret. Valid questionnaires are available in the marital area for assessing communication (e.g., Navran, 1967) and satisfaction (e.g., Locke & Wallace, 1959). For child treatment studies, there are numerous teacher and parent rating scales that have demonstrated reliability and validity (Achenbach, 1978; K. D. O'Leary & Johnson, in press; Quay, in press). On the other hand, valid self-report questionnaires that assess self-esteem and attitudes of children toward parents and school are needed (cf. Cooper-smith, 1967).

Even though all of the problems associated with self-report measures cannot be elimi-

nated, we strongly recommend their use because they are focused, convenient, and economical. Further, certain areas cannot be assessed by observational measures (e.g., perception of feelings toward one's spouse, expectations regarding marriage, attitudes of children toward teachers, and self-esteem).

Observational Measures

The problems discussed above in regard to self-report measures indicate a clear need for direct observation of the behaviors of interest. Direct, or naturalistic, observation has become a hallmark of research in behavior therapy, and certain journals will often not accept manuscripts that do not include such data (e.g., *Journal of Applied Behavior Analysis*). Direct observations usually provide less biased data than self-reports, and they can be used to obtain information not easily provided by clients (e.g., data on facial expressions or other nonverbal aspects of communication of which a spouse or parent may not be aware). Finally, the collection of audiotapes and/or videotapes allows for the possible reanalysis of data by different investigators who wish to compare and contrast treatments.

In the child treatment area, observational coding systems have been used extensively in classrooms (Kent & Foster, 1977). Classroom observations have proven reliable, valid, sensitive to treatment changes; are largely non-reactive (S. G. O'Leary & K. D. O'Leary, 1976); and are capable of distinguishing control from clinical populations (Abikoff, Gittelman-Klein, & Klein, 1977). In homes or simulated home settings, direct observations of children have also proven reliable, valid, and sensitive to treatment changes (Patterson, 1974), but sometimes they are reactive; that is, the presence of an observer in a home can alter the nature of parent-child interactions (White, 1977). The problem of reactivity is an important one that is often overlooked and may be especially problematic with adolescents when only a few observations are made (K. D. O'Leary & Johnson, in press).

Marital interaction codes have been developed that discriminate between distressed and nondistressed couples (Rausch, Barry, Hertel, & Swain, 1974; Gottman, Notarius, & Mark-

man, Note 3). However, with few exceptions, complex coding systems have not been used as pre-post dependent measures (e.g., Weiss, Hops, & Patterson, 1973). The use of such observations is strongly recommended, and several methodological issues should be considered. In marital treatment research the observational context is generally a laboratory or interview setting. Many investigators use variations of Strodbeck's revealed difference technique (Strodbeck, 1951), in which the experimenter identifies differences of opinion between the spouses and then asks them to resolve these differences and reach mutually agreed-on conclusions. Typically, the couple is left alone when discussing the conflicts, and their interaction is videotaped or audiotaped. When using this procedure, discussions may occur that evoke very hostile feelings. Thus, it may be ill-advised to assess certain populations (e.g., physically violent couples) in this manner. In all cases, the investigator should conduct at least a brief clinical interview following the task to insure that the spouses do not leave the assessment interview angrier with each other than when they arrived.

The content of the conflicts discussed may have a critical influence on the data obtained. For example, some evidence indicates that with older couples a high degree of induced conflict is necessary to discriminate the communication patterns of distressed and nondistressed couples (Gottman, Notarius, Gonso, & Markman, 1976). In the Gottman et al. study, the difference between the high and low conflict situations involved the degree to which the topic discussed was relevant to the marital relationship. One implication of this finding is that topics relevant to the couples' problems produce more realistic samples of communication.

A method used to insure a highly relevant discussion has been to ask the couples to discuss their own marital problems. However, we feel that this procedure has a serious methodological problem in that a change in the interactions may reflect either an amelioration in the particular presenting complaints or an alteration in the general communication style of the couple. Presumably, general

changes in communication and problem-solving skills should be measured on topics relevant to, but not identical with, the couple's presenting problems if predictions are to be made about how the couple will handle conflicts that arise in the future.

One additional methodological issue in the marital area involves selecting the kinds of behaviors to observe and the means to assess them. Both the affective and content domains of communication should be assessed to discriminate between distressed and nondistressed couples (Gottman et al., 1976). Although it is not clear that videotapes offer incremental validity over audiotapes, the affective quality of an exchange ought to be more easily assessed on videotapes (e.g., sneers, eye contact, body position). However, since video recordings are costly, potentially reactive, and unavailable in most nonresearch clinics, reliable, discriminating codes for rating audiotapes should be developed.

At present, observational methodology in the child treatment area, particularly in classrooms, is more sophisticated than that in marital therapy research. Further development of behavioral codes of marital interactions and parent-child interactions in the home is mandated. Observational methodology is costly and time-consuming, but obtaining independent observations that supplement self-report data is usually crucial.

Treatment Specification

Therapy Manuals

Failure to specify therapeutic procedures in detail is one of the most common and yet most serious problems in psychotherapy research. This error is largely an error of omission, and to rectify the problem some journal editors (e.g., *Cognitive Therapy and Research*, *Journal of Applied Behavior Analysis*) have recently required that treatment manuals be made available for interested readers through the author, the National Auxiliary Publications Service, or some other central information source. Without such manuals or other training materials (e.g., video or audio cassettes), replication of treatment studies is almost impossible.

Training Programs

In addition to the description of the actual treatment, a manuscript should include the following information on the therapists' preparation: the length and type of training, that is, pilot cases, role playing, reviewing tapes, and books, manuals, and other reading materials used. The amount and frequency of therapy supervision should also be specified. Since complete specification of the training program would be impossible in most journals, details regarding the training programs should be explicated in therapy manuals.

Assessment of the Independent Variable(s)

To state unequivocally that a treatment is effective, some measure that the treatment was actually implemented must be obtained. In psychological research, several levels of analysis can be used to assess the independent variable. On one level, an investigator can determine whether the therapists acted in accord with the therapeutic regimen. At a second level, with regard to certain types of therapy (e.g., behavior therapy), a determination can be made of whether the clients actually implemented the procedures recommended by the therapists. For example, in a communication program for parents and children, one should assess whether the therapists gave the appropriate suggestions as outlined in the therapy manual. It can also be determined whether or not the parents followed the designated communication exercises with their children at home.

Checks on the independent variables in treatment research with children and marital partners are seldom reported or are incomplete. Both therapists' and clients' self-reports, as well as observations of therapy programs or analyses of therapy tapes, are strongly advocated. Having a detailed behavioral assessment of the independent treatment variables prompts investigators to provide adequate treatment descriptions, including the percentage of time therapists spent in various treatment activities. Most importantly, when comparisons between different treatment programs are made, an investigator should provide reliable, objective data to document the

actual procedural differences that distinguished the various treatment groups.

Experimental Design

Aspects of therapist-client interactions that should be controlled for include attention to a client's problem, expectations about therapeutic outcome, and other nonspecific relationship factors. Traditionally, solutions for such problems have been placebo or attentional control groups. We consider placebo groups to be ethically and methodologically problematic, and a number of alternatives have been suggested to control for therapist attention and client expectations (K. D. O'Leary & Borkovec, in press). A few of these choices are (a) the use of the best available alternatives (Jacobson & Baucom, Note 4), that is, comparing one form of treatment with another commonly accepted treatment for a particular problem; (b) component control comparisons in which procedural elements from a total treatment package are evaluated; (c) evaluation of the same treatment program under two different conditions—a neutral expectancy control group that receives a neutral expectation set and a second group that receives usual or exaggerated positive expectation sets; and (d) counterdemand manipulations in which clients are told that they should not expect improvement until after a fixed number of sessions. A full discussion of controls for expectations, therapist attention, and relationship factors is clearly beyond the scope of this article, but such factors must be addressed by any serious researcher.

For problems that change with maturation or remit spontaneously, waiting list controls for limited periods may be especially useful. Clearly, knowing whether an intervention is more effective than none at all is absolutely necessary. In child treatment evaluations, maturational controls are critical (Kent, 1976), and although many people would think that marital cases are generally unlikely to change during short periods without treatment (Gurman, 1975), the inclusion of waiting list control groups can result in tempered claims regarding the effectiveness of an intervention (K. D. O'Leary & Turkewitz, 1978). When ethical concerns do not allow for inclusion of a standard waiting list control

group, the clients on the waiting list can be monitored, and those who evidence deterioration can be placed in the treatment group (Stuart, 1973).

When comparing different treatment programs, some assessment should be made regarding the clients' acceptance of the therapeutic rationale (Kazdin & Wilcoxon, 1976). In addition, clients should be asked to evaluate the therapists as well as the therapeutic procedures (Kazdin, 1977; Kent & O'Leary, 1976; Wolf, 1978).

Single-subject or within-subject designs have been used by both child and marital researchers, and they have been especially useful in evaluating the effectiveness of treatment procedures when (a) the effect can ethically be reversed (reversal designs) or (b) generalization of effects across behaviors is small (multiple baseline designs). When either of these conditions is unlikely, group designs should be used. (For a full discussion of single subject methodology, consult Hersen & Barlow, 1976.)

In addition to deciding on adequate control groups and research designs, plans should be made to collect follow-up data to determine whether the treatment effects persist when the natural environment is not reprogrammed by the therapist. An assessment of the posttreatment environment is advisable—especially with children—to ascertain the conditions under which behavior is or is not maintained (Bijou, 1974).

Data Analysis and Interpretation

A common error in analyzing data from dyadic interactions in the child and marital areas is to include both partners from the dyad in the degrees of freedom. Since the members of the dyad influence one another, they should not be regarded as independent units in analyses of variance. For example, in examining the simple pre-post effects of an intervention on 10 couples, analyses using data from individual subjects should be partitioned as follows: pre-post, male versus female, the interaction between male-female and pre-post, and couple differences. Alternatively, the average of the husband's and wife's scores could be treated as a single unit.

If a therapist conducts group therapy, as

is often the case in marital research (Gurman & Kniskern, 1978), generalizations about therapeutic effects must be limited to a group context and the number and types of therapist(s) used. Further, the nonindependence of subjects must be considered in determining the degrees of freedom for testing group effects. As a general rule, the number of therapy groups in each experimental condition should be large enough to detect differences that cannot be attributed to idiosyncratic group characteristics.

If several dependent measures are used and the correlations between the measures are high, a multivariate analysis of variance that takes the intercorrelations into account should be considered (C. L. Olson, 1976).

Practicing clinicians who are the major consumers of clinical research are concerned about individual variability (range and standard deviation) in response to treatment. Consequently, estimates of the strength of association (omega-squared) should be reported wherever possible to allow the clinician to ascertain the impact of the treatment (Hays, 1963). That is, it is useful to know the percentage of variance in a dependent measure accounted for by the treatment.

Subject attrition and incomplete data can present serious problems in interpretation. In a review of psychotherapy literature, Baekeland and Lundwall (1975) found that clients who dropped out of treatment were more likely to be from lower socioeconomic groups and to have little affiliation with others. As such factors may relate to outcome, it is essential to report the percentage of dropouts. Additionally, one should attempt to obtain posttermination data from dropouts; we have found that some clients are willing to complete the assessment even if they are unwilling to continue in therapy. There are occasions when clients participate in the entire therapy program but do not complete the posttreatment or follow-up assessment materials. In an evaluation of a child clinic, Olmanns, Broderick, and O'Leary (1977) found a significant relationship between difficulty in obtaining outcome data and responsiveness to treatment. Thus, both therapy dropouts and missing data points should be considered as potential sources of positive bias.

Conclusion

Salient methodological errors in conducting child and marital treatment research include inadequate client and therapist selection and description, unsubstantiated labels and diagnoses, lack of multiple outcome criteria, use of dependent measures of unspecified reliability and validity, and failure to provide treatment manuals and to check whether the specific treatments were in fact implemented. Experimental design issues are complex and relate to the substantive questions of interest. Very common problems, however, involve the need to control for expectation, therapist attention, and relationship factors, and the failure to conduct follow-up evaluations. Finally, common errors in data analysis and interpretation have been noted. Throughout the text suggestions have been made to help minimize the methodological errors discussed.

Reference Notes

1. Turkewitz, H. T., & O'Leary, K. D. *Communication and behavioral marital therapy: An outcome study*. Paper presented at the Association for Advancement of Behavior Therapy workshop, New York, December 1976.
2. O'Leary, S. Personal communication, September 10, 1977.
3. Gottman, J., Notarius, C., & Markman, H. *Couples Interaction Scoring System (CISS)*. Unpublished manuscript, Indiana University, 1976.
4. Jacobson, N. S., & Baucom, D. H. *Design and assessment of nonspecific control groups in behavior modification*. Paper presented at the meeting of the Association for Advancement of Behavior Therapy, San Francisco, December 1975.

References

- Abikoff, H., Gittelman-Klein, R., & Klein, D. F. Validation of a classroom observation code for hyperactive children. *Journal of Consulting and Clinical Psychology*, 1977, 45, 772-783.
- Achenbach, T. M. Psychopathology of childhood: Research problems and issues. *Journal of Consulting and Clinical Psychology*, 1978, 46, 759-776.
- American Psychological Association. *Publication manual of the American Psychological Association* (2nd ed.). Washington, D.C.: Author, 1974.
- Baekeland, F., & Lundwall, L. Dropping out of treatment: A critical review. *Psychological Bulletin*, 1975, 82, 738-783.
- Bienvenu, M. J. Measurement of marital communication. *Family Coordinator*, 1970, 19, 26-31.

- Bijou, S. W. Focus. *Journal of Applied Behavior Analysis*, 1974, 7, inside back cover.
- Cerreto, M. C., & Tuma, J. M. Distribution of DSM-II diagnoses in a child psychiatric setting. *Journal of Abnormal Child Psychology*, 1977, 5, 147-156.
- Chess, S., & Thomas, A. *Annual progress in child psychiatry and child development*. New York: Brunner-Mazel, 1977.
- Connors, C. K. A teacher rating scale for use in drug studies with children. *American Journal of Child Psychiatry*, 1969, 126, 884-888.
- Coopersmith, S. *The antecedents of self-esteem*. San Francisco: Freeman, 1967.
- Edmonds, V. H. Marital conventionalization: Definition and measurement. *Journal of Marriage and the Family*, 1967, 29, 681-688.
- Fry, E. B. A readability formula that saves time. *Journal of Reading*, 1968, 11, 513-516.
- Gottman, J., Notarius, C., Gonso, J., & Markman, H. *A couple's guide to communication*. Champaign, Ill.: Research Press, 1976.
- Gurman, A. S. The effects and effectiveness of marital therapy. In A. S. Gurman & D. G. Rice (Eds.), *Couples in conflict*. New York: Aronson, 1975.
- Gurman, A. S., & Kniskern, D. P. Research on marital and family therapy: Progress, prospective and prospect. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change*. New York: Wiley, 1978.
- Harrell, J., & Guernsey, B. G. Training married couples in conflict negotiation skills. In D. H. L. Olson (Ed.), *Treating relationships*. Lake Mills, Iowa: Graphic, 1976.
- Hartmann, D. P., Roper, B. L., & Gelfand, D. M. An evaluation of alternative modes of child psychotherapy. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology*. New York: Plenum Press, 1977.
- Hays, W. L. *Statistics for psychologists*. New York: Holt, 1963.
- Hersen, M., & Barlow, D. H. *Single case experimental designs*. Elmsford, N.Y.: Pergamon Press, 1976.
- Johnson, S. M., & Christensen, A. Multiple criteria follow-up of behavior modification with families. *Journal of Abnormal Child Psychology*, 1975, 3, 135-154.
- Kazdin, A. E. Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, 1977, 1, 427-452.
- Kazdin, A. E., & Wilcoxon, L. Systematic desensitization and nonspecific treatment effects: A methodological evaluation. *Psychological Bulletin*, 1976, 83, 729-758.
- Kenny, T., et al. Characteristics of children referred for hyperactivity. *Journal of Pediatrics*, 1971, 79, 618-622.
- Kent, R. N. A methodological critique of "Interventions for boys with conduct problems." *Journal of Consulting and Clinical Psychology*, 1976, 44, 297-299.
- Kent, R. N., & Foster, S. Direct observational procedures: Methodological issues in naturalistic settings. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Kent, R. N., & O'Leary, K. D. A controlled evaluation of behavior modification with conduct problem children. *Journal of Consulting and Clinical Psychology*, 1976, 44, 586-596.
- Kimmel, D., & Van der Veen, F. Factors of marital adjustment in Locke's Marital Adjustment Test. *Journal of Marriage and the Family*, 1974, 36, 57-63.
- Lang, P. J. The application of psychophysiological methods to the study of psychotherapy and behavior modification. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change*. New York: Wiley, 1971.
- Liberman, R. P., Levine, J., Wheeler, E., Sanders, N., & Wallace, C. J. Marital therapy in groups. A comparative evaluation of behavioral and interactional formats. *Acta Psychiatrica Scandinavica*, 1976 (Suppl. 266), 1-34.
- Locke, H. J., & Wallace, K. M. Short marital-adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living*, 1959, 21, 251-255.
- LoPiccolo, J. Methodological issues in research on treatment of sexual dysfunction. In R. Green & J. Wiener (Eds.), *Methodological issues in sex research*. Washington, D.C.: U.S. Government Printing Office, 1978.
- Murstein, B. I., & Beck, G. D. Person perception, marriage adjustment, and social desirability. *Journal of Consulting and Clinical Psychology*, 1972, 39, 396-403.
- Navran, L. Communication and adjustment in marriage. *Family Process*, 1967, 6, 173-184.
- O'Leary, K. D., & Borkovec, T. D. Placebo groups: Unrealistic and unethical controls in psychotherapy research. *American Psychologist*, in press.
- O'Leary, K. D., & Johnson, S. B. Assessment of psychopathology in children. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of children* (2nd ed.). New York: Wiley, in press.
- O'Leary, K. D., & Turkewitz, H. T. Marital therapy from a behavioral perspective. In T. Paolino & B. McCrady (Eds.), *Marriage and marital therapy: Psychoanalytic, behavioral and systems theory perspectives*. New York: Brunner-Mazel, 1978.
- O'Leary, K. D., & Wilson, G. T. *Behavior therapy: Application and outcome*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.
- O'Leary, S. G., & O'Leary, K. D. Behavior modification in the school. In H. Leitenberg (Ed.), *Handbook of behavior modification and behavior therapy*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.
- Olson, C. L. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 1976, 83, 579-586.
- Olson, D. H. Marital and family therapy: Integrative review and critique. *Journal of Marriage and the Family*, 1970, 32, 501-538.

- Oltmanns, T. F., Broderick, J. E., & O'Leary, K. D. Marital adjustment and the efficacy of behavior therapy with children. *Journal of Consulting and Clinical Psychology*, 1977, 45, 724-729.
- Patterson, G. R. Interventions for boys with conduct problems: Multiple settings, treatments and criteria. *Journal of Consulting and Clinical Psychology*, 1974, 42, 471-481.
- Quay, H. C. Classification: Patterns of aggression, anxiety-withdrawal, and immaturity. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood* (2nd ed.). New York: Wiley, in press.
- Rausch, H. L., Barry, W. A., Hertel, R. K., & Swain, M. A. *Communication, conflict, and marriage*. San Francisco: Jossey-Bass, 1974.
- Ross, S., & Ross, D. *Hyperactivity*. New York: Wiley, 1977.
- Routh, D. K., Schroeder, C. S., & Otuama, L. A. Development of activity level in children. *Developmental Psychology*, 1974, 10, 163-168.
- Strodtbeck, F. L. Husband-wife interaction over revealed differences. *American Sociological Review*, 1951, 16, 468-473.
- Strupp, H. H., & Hadley, S. W. A tripartite model of mental health and therapeutic outcomes: With special reference to negative effects in psychotherapy. *American Psychologist*, 1977, 32, 187-196.
- Stuart, R. B. Notes on the ethics of behavior research and intervention. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice*. Champaign, Ill.: Research Press, 1973.
- Weiss, R. L., Hops, H., & Patterson, G. R. A framework for conceptualizing marital conflict: A technology for altering it, some data for evaluating it. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice*. Champaign, Ill.: Research Press, 1973.
- Wells, R. A., Dilkes, T. C., & Burckhardt, N. T. The results of family therapy: A critical review of the literature. In D. H. L. Olson (Ed.), *Treating relationships*. Lake Mills, Iowa: Graphic Press, 1976.
- White, G. D. Effects of observer presence on family interaction. *Journal of Applied Behavior Analysis*, 1977, 10, 734.
- Williams, W. S. Class differences in the attitudes of psychiatric patients. *Social Problems*, 1956, 4, 240-244.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.
- Wolf, M. M. Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, 1978, 11, 203-215.
- Wolman, B. B., Egan, R. A., & Ross, A. O. *Handbook of treatment of mental disorders in childhood and adolescence*. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Received December 12, 1977 ■

Psychopathology of Childhood: Research Problems and Issues

Thomas M. Achenbach
National Institute of Mental Health
Bethesda, Maryland

It is argued that research on child psychopathology would benefit from reducing the influence of adult treatment models and from applying a developmental perspective to clinical research on children. Specific methodological problems are discussed, including the assessment of subject characteristics; the use of replicable and generalizable diagnostic classifications; the effects of situational specificity and developmental variance on measures of children's behavior; the need to avoid pathological biases in judging children; relationships among correlation, causation, and prediction in a developmental context; problems in measuring change; and the effects of age, cohort, and time of measurement, as well as fallacies in drawing longitudinal conclusions from cross-sectional data. Outstanding research needs are also identified, including the need to devise and use well-standardized measures; the need to evaluate interactions between subject and treatment variables in outcome research; the need for long-term follow-ups of children identified as being at risk; the need for cumulative programmatic research; and the need to link research more closely to service systems.

Research on child psychopathology has traditionally been treated like a stepchild for whom no one assumes full responsibility. Despite clinical emphasis on the childhood roots of adult disorders, psychopathology has been studied far more intensively in adults than in children. Furthermore, research on psychopathology in children has been unduly shaped by a mental health system in which clients must adopt a patientlike role. This system may not be entirely unjustified for adults, but it engenders assumptions that are certainly inappropriate for children.

First, children rarely have realistic conceptions of psychopathology and mental health services. Second, both the judgment that a child needs help and the initiative required

to obtain it originate with adults rather than with the child. Third, children have much less freedom to alter their circumstances than do adults, and adult-centered views of treatment as primarily a relationship between therapist and patient neglect children's overwhelming dependence on their families. Fourth, unlike adults who have reached plateaus in their social, educational, cognitive, and physical growth, children must be viewed in relation to their progress along these dimensions. Rather than judging clients' problems in terms of interference with current capabilities—as may be appropriate with adults—it is necessary to judge children's problems in terms of their interference with future development. Unless children can progress in knowledge, skills, and ways of relating to others, they are unlikely to achieve a successful adaptation in the long run, no matter how satisfactory an outcome may appear in the short run.

To help troubled children, we need a far greater investment of resources, energy, and brain power in objectively assessing their needs and in developing and evaluating sys-

The author is grateful to Craig Edelbrock, Gale Inoff, and John Weisz for their critical readings of the manuscript.

Requests for reprints should be sent to Thomas M. Achenbach, Laboratory of Developmental Psychology, Building 15K, National Institute of Mental Health, 9000 Rockville Pike, Bethesda, Maryland 20014.

tematic approaches to meeting these needs. Despite a flood of rhetoric and legislation on behalf of children, there is little evidence of cumulative progress in alleviating children's mental health problems. In my view, it is the responsibility of mental health professionals not only to advocate funds and programs for children but to begin demonstrating more accountability for the services rendered. The lack of evidence for the efficacy of most mental health services to children (cf. Achenbach, 1974; Levitt, 1971; Shepherd, Oppenheim, & Mitchell, 1971) behooves mental health professionals to assume greater initiative in determining how disturbed children can in fact be helped. Reform is vitally needed to improve the mental health services available to children. However, meaningful reforms can only originate from within the mental health professions through better validation and use of whatever is good in the present system and development of new approaches to replace whatever cannot be firmly validated.

In hope of bridging the gap between the pressing needs of the mental health service system and the scientific principles to which research must adhere, it may be helpful first to consider some possible reasons for the dearth of high-quality research on behalf of disturbed children. Thereafter, I will deal with specific methodological problems that arise in research on child psychopathology. I will conclude with what I consider to be some of the most important needs for research in this area.

Reducing Some Obstacles to Research on Child Psychopathology

The stepchild status of research on child psychopathology results, I think, from the fact that most clinical training is oriented primarily toward treatment of adults. Those who wish to work with children must typically defer child clinical experience until after they have a heavy dose of adult training. Even though adult training may aid in dealing with the parents of disturbed children, it fosters treatment models and attitudes that are inappropriate not only for treating children but also for developing researchable hypotheses about psychopathology in children. The em-

phasis on treatment of adults stems no doubt from the fact that adults have long been the primary paying customers for mental health services. An additional factor may be that treatment of adults offers a glamorous and prestigious role more in keeping with the professional status and personal insights sought by those who enter clinical training. As a result, there has been insufficient incentive to develop an independent tradition of child clinical training and research. Instead, most research on child behavior has been left to academic developmental psychologists who are not attuned to the functioning and needs of the mental health service system. Widening the gulf still further between the "two cultures" of researchers and practitioners is the current preoccupation of clinical psychologists with winning recognition as independent service providers.

Several changes of orientation might help to overcome these obstacles to research on child psychopathology. An especially desirable change would be to make child clinical training a primary specialty rather than allowing it to remain secondary to adult training. This could be done in a number of ways. One way would be to make child training a separate and distinct track coequal with adult training in the larger clinical programs. A second way would be to give more emphasis to the differences between approaches appropriate to children and adults in general clinical training programs, supplemented with instruction in normal child development. A third way would be to encourage more training programs to make child clinical their primary and official emphasis, rather than continuing to have so many child practitioners trained in general clinical programs geared almost exclusively to treatment of adults.

Another desirable change would be in career roles for psychologists whose talents lie in research. A major portion of biomedical research is carried out by full-time researchers who function as members of teams or laboratories specializing in specific organic diseases and abnormalities. By contrast, the dominant model for researchers in psychopathology has been that of the teacher cum researcher cum clinician. Attempts to simultaneously fill all three roles may have interfered with pro-

grammatic research in the best of times, but shrinking resources, the disappearance of teaching positions, and increasing demands on those who do teach mean that alternatives to the primarily academic and primarily service career models must be fostered if behavioral research on psychopathology is to progress.

A further desirable change would be to link the study of psychopathology in children more closely to the study of normal development. This would entail delineating the specific ways in which disorders disrupt the typical course of cognitive, emotional, behavioral, and social development and how the outcomes of various disorders affect long-term adaptation. When considered in a developmental context, some disorders that seem ominous may not in fact be debilitating unless they prevent children from receiving socialization experiences necessary for adaptation to their culture. Other disorders that seem less ominous might have severe consequences if they interfere with the cumulative socialization needed for long-term adaptation. A developmental perspective on child psychopathology would emphasize (a) the transactional relationships between children's characteristics as individuals and the social contexts in which they function and (b) the different implications that various combinations of child and environment have at different developmental levels. Although linking childhood characteristics with adult psychopathology should remain an important goal, research on child psychopathology would benefit from greater reliance on knowledge of normal development than on models of adult psychopathology.

Methodological Problems

The emphasis here will be on problems evident in manuscripts submitted to the *Journal of Consulting and Clinical Psychology*, but they are representative of weaknesses in work appearing elsewhere as well. Although some of the problems are quite general, I will discuss them as they arise in the context of research on child psychopathology.

Subject Characteristics

The choice of subjects in too many studies is determined more by convenience than by

the requirements of sound research design. The design of every study should take careful account of such variables as the developmental level, intelligence, sex, race, socioeconomic status (SES), and clinical status of subjects.

Developmental level. Even small differences in developmental level can have large effects on subjects' capabilities, the ways in which they construe situations, the kinds of experiences they have had, and the behavior they elicit from others. Chronological age (CA) is not an adequate index of developmental level if other relevant aspects of development—such as cognitive ability, physical maturation, academic achievement, or level of schooling—deviate from CA norms for the population to which findings are to be generalized. If intellectual ability deviates from these norms, an index of cognitive level should be obtained, such as mental age (MA) or total raw score on an IQ test. It should be remembered that the IQ is not a measure of cognitive level but of the deviation of an individual's cognitive level from that of a normative group of comparable CA. Matching subjects for IQ or partialing IQ out of relationships among other variables thus controls for differences in cognitive level only if the subjects are of exactly the same CA. Since this is rarely the case, MA is more appropriate than IQ for use in equating subjects for cognitive level, although the choice of a measure for cognitive level should be guided by the questions that the research is to answer. For example, the Performance scale of the Wechsler Intelligence Scale for Children or nonverbal Piagetian measures might be necessary for use with language-deficient subjects, whereas the Binet MA might be more satisfactory for other groups.

If groups with IQs outside the average range are studied, behavioral expectations should be based as much on their cognitive levels as on their CA. A frequent error is to compare the behavior of children having normal IQs to that of retarded children of the same CA and then to interpret differences as reflecting defects inherent in mental retardation. Since retarded children—by definition—have lower MAs than normal children of the same CA, attributing differences to retarda-

tion per se is tautological at best. Worse than being tautological, however, it may be patently wrong to attribute behavioral differences to the IQ differences, because developing more slowly than the norm for one's culture has many consequences besides the effects of whatever causes the slow development. These consequences include stigmatic labels, a high rate of failure when CA-appropriate tasks are attempted, ridicule, deprivation of positive adult attention, and atypical educational regimens. All of these factors may affect learning, personality, and motivational variables not directly affected by slow cognitive development per se (cf. Zigler, 1971).

Researchers must also be sensitive to the possible effects of cognitive level and atypical rates of development in children who do not happen to be called retarded. For example, characteristics other than low IQ earn children labels such as *psychotic*, *delinquent*, or *disadvantaged*. However, the cognitive functioning of such children is often below the norm for their CA. When comparisons with normal children reveal inferiorities, they should not be attributed to psychosis, delinquency, or environmental disadvantages unless the effects of cognitive level and the concomitants of slow intellectual development can be ruled out through comparisons with children matched for cognitive level (e.g., retarded children and/or younger normal children).

Demographic variables. Experiential differences related to sex, race, and SES are as likely to affect behavior as are the experiential differences related to cognitive level. Demographic characteristics of child subjects may also affect the behavior of adult experimenters toward the subjects and vice versa, resulting in interactions between the effects of child and experimenter characteristics (Back & Dana, 1977; Marwit & Neumann, 1974; Quereschi, 1968).

Despite the influence of sex, race, and SES differences, some otherwise high-quality studies describe their samples only as "children," with no mention at all of these variables, much less any indication that their effects were tested or controlled for (e.g., Kent & O'Leary, 1976). In other studies, the analyses of demographic variables are confounded

in such a way as to obscure possible interactions among them (e.g., Gesten, 1976). Although limitations on subject pools and the different incidence of particular disorders in various groups often preclude exhaustive testing of the effects of demographic variables, their distributions within subject samples should at least be described in detail, and analyses involving them should be sensitive to possible interactions. When the effects of such variables cannot be tested, investigators should explicitly note that their findings are limited to populations like those represented by their samples.

Sources of subjects. The settings from which children are obtained can exert major influences on behavior. The different experiential histories, motivational structures, and expectations of subjects obtained through public schools, clinics, residential institutions, and courts can produce major differences in behavior despite similarities in developmental level, demographic characteristics, and diagnosis. It has been amply demonstrated, for example, that the deprivation of adult attention experienced by institutionalized children can affect their behavior on experimental tasks (cf. Zigler, 1971). In addition, because children do not always distinguish among occupational roles within a setting, their attitudes toward familiar adults are likely to generalize to research personnel. It is therefore essential that sources of subjects be clearly described and that the setting be treated as an important variable that can affect results.

Although control over subject and setting variables is usually sought through standardization of instructions and procedures, the variety of settings in which disturbed children typically must be studied may preclude standardization of all relevant variables. Furthermore, the fragmentation of services for children can create selective biases in samples obtained from each setting. Because subject, setting, and sampling variables can seldom be fully controlled, an alternative is to replicate findings with subjects obtained from a variety of settings. Findings that are robust despite differences in settings have the greatest breadth of application, but it is also important to identify the limitations that specific settings impose on findings.

Diagnostic Classification

Perennial controversies over diagnosis reflect the immaturity of psychopathology as a field of study. Even more than adult psychopathology, child psychopathology lacks a coherent conceptual framework for describing and discriminating among disorders, much less for supporting inferences about the etiology, course, or appropriate treatments for specific disorders. Not only is there no accepted diagnostic system, but there are fundamental disagreements about whether diagnosis is a legitimate enterprise at all (cf. Hobbs, 1975). Critics justly contend that diagnostic labels may stigmatize children without gaining them the benefits of appropriate services. Researchers must be sensitive to the whole child and combat indiscriminate summary labels for individual children. Yet there is no way to accumulate and transmit knowledge without conceptual categories. Findings cannot be generalized unless researchers group and describe subjects according to categories that can be reliably used by others.

In the absence of a generally accepted taxonomy, diagnostic classification schemes must be chosen to fit particular populations and research aims. The portrayal of childhood disorders in the official system for clinical records, embodied in the second edition of the American Psychiatric Association's (1968) *Diagnostic and Statistical Manual of Mental Disorders* (DSM II), represents an advance over the first edition of the DSM (American Psychiatric Association, 1952) in being more differentiated and based at least partly on empirical studies. However, no operational definitions are provided for its categories, and the categories do not reflect developmental differences. Furthermore, research indicates that interjudge reliability averages only about 60% for broad categories, such as psychotic versus neurotic, and considerably less for specific subcategories (cf. Achenbach & Edelbrock, in press). The forthcoming DSM III is intended to remedy at least some of these deficiencies, but current drafts inspire no hope of a panacea.

An alternative system has been proposed by the Group for the Advancement of Psychiatry (1966). It is designed specifically for

childhood disorders, is far more differentiated than the DSM II, and reflects developmental considerations. However, like the DSM, its categories consist of narrative mixtures of description and inference formulated by committees of psychiatrists rather than being operationally defined on the basis of empirical research. Interjudge reliability is no better than for the adult categories of the DSM (Freeman, 1971).

A third alternative is to use categories derived through multivariate analyses of behavior checklists. Despite differences in checklists, subject populations, types of raters, and methods of analysis, there is considerable convergence in the behavioral syndromes identified in various studies (cf. Achenbach & Edelbrock, in press). Much remains to be done by way of translating these syndromes into categories of individuals or disorders, but several of the checklists have been sufficiently validated to provide a basis for describing and categorizing disturbed children for research purposes. Even if investigators prefer a different approach to diagnostic classification, the generalizability of their work will be greatly enhanced by scoring subjects on at least one well-validated checklist.

In addition to identifying behavioral characteristics as objectively as possible, it is important to identify any organic conditions and medications that may be relevant. Even in studies focused on nonorganic characteristics, differences in the distribution of organic conditions among subject groups can affect findings. This is just as true for disorders in which organic etiologies have not been proven as for disorders in which organic etiologies are known. As inconclusive as organic diagnoses may be for a particular study, they almost always have potential implications for the findings. For example, in a comparison of black children and white children, it was found that low-IQ black subjects obtained higher Rorschach perceptual integration scores than did low-IQ white subjects (Gerstein, Brodzinsky, & Reiskind, 1976). This was interpreted as indicating that standard IQ tests may underestimate the cognitive capacities of black children. However, the only diagnostic information given was that "all groups were heterogeneous with respect to diagnostic cate-

gory" (Gerstein et al., p. 761). A low-IQ clinic population is likely to contain children with neurological dysfunctions, and, as Gerstein et al. noted, such dysfunctions have been found to be related to poor integrative performance on the Rorschach. It is therefore possible that differences in the distribution of neurological dysfunctions in the black samples and white samples could account for the racial differences in disparities observed between IQ and Rorschach performance. The value of the study could have been greatly enhanced if organic dysfunctions in the samples had been controlled for, their effects tested, or at least reported.

Measures of Children's Behavior

A great many measures of children's behavior have been devised, but few have been adequately standardized and validated for studying psychopathology. Because of the time, sample sizes, and effort required for standardization, investigators are often forced to choose between devising a new measure without adequate standardization or employing a measure that was standardized for populations or uses not entirely appropriate for the investigator's purposes. The result is both a proliferation of measures having unknown properties and an accumulation of apparent failures of standardized measures to deliver what they promised. The latter problem may be due in part to the failure of authors of standardized measures to emphasize the limitations on the populations and situations for which their measures are appropriate. However, because authors cannot foresee all possible uses and abuses of their measures, the measures should not be blamed if they fail to fulfill purposes for which they were not validated. Careful attempts to broaden the application of a measure are warranted as long as poor results are not interpreted to mean that the measure is invalid for its intended purpose. But, lacking an armamentarium of well-validated research instruments, we have a collective responsibility to insure that existing instruments are properly used rather than unfairly undermined through inappropriate applications.

Whether measures have previously been

standardized or are constructed for a particular study, it is always necessary to pilot test them with subjects and conditions like those to be studied in the research. The variability in children's expectations about research situations, their attentional and cognitive limitations, and other variables that affect their responses to research procedures make it essential to obtain a child's eye view of the research situation in order to eliminate sources of anxiety, misunderstanding, response sets, and demand characteristics (cf. Achenbach, 1978b, chap. 7). Careful observation and interviewing of pilot subjects are therefore necessary precursors of all research with children. Once procedures have been finalized, they should be documented and reported in sufficient detail to enable others to replicate them with no inadvertent variations that might affect findings.

Situational specificity of behavior. Whatever position one takes in the debate over the relative dominance of person or situation variance in behavior (cf. Mischel, 1977), it is obvious that children's behavior varies tremendously with the situation. In ratings of normal nursery school children, for example, Rose, Blank, and Spalter (1975) found very little consistency from one situation to another even within a nursery school environment, despite high reliability among observers. This does not necessarily mean that no stable subject variance was detectable, as significant correlations were found between reratings of children in the same *situations* within the nursery school 4 months apart.

In addition to situational effects, the type of observer and his/her relationship to child subjects inevitably affect the behavior reported. The greater the disparity between the situations in which observers view children and the more the observers differ in their relationships to the children, the lower the agreement in observations (cf. Achenbach & Edelbrock, in press). Because there is typically no single criterion situation against which to validate observations, it is important to obtain multiple measures from observers who differ in their relationships with the subjects. Despite their potential biases, parents are usually the key informants regarding behavior of clinical concern, because their per-

ceptions determine what will be done about their children. Furthermore, parents' reports of behavior problems have been found to be much more complete than those of teachers, school observers, home observers, or clinic intake workers (Novick, Rosenfeld, Bloch, & Dawson, 1966). Interparent agreement and the agreement of teachers and clinicians with parents have also been found to exceed agreement between clinicians, as well as agreement between clinicians and teachers (Miller, 1964). Furthermore, clinical trainees' judgments of children's pathology have been found to be influenced more by parents' reports than by direct observations of children in a clinical setting (McCoy, 1976). Whenever possible, studies of child psychopathology should therefore take parent reports into account.

In addition to parents, alternative sources include other family members, teachers, clinicians, peers, and self-reports. The weight given to reports from these sources should depend on the cross-validated accuracy with which they are found to relate to the variables of interest. For example, peer reports may be among the best predictors of behavior in group social situations, teacher reports may be the best predictors of academic functioning, and clinician reports may be the best predictors of behavior during psychotherapy sessions. When reports from various sources disagree, choices must be made about the relative importance of situations to which findings are to be applied.

Developmental versus trait variance. An important goal of most behavioral research is to identify relationships among various measures of behavior. This goal has traditionally been most prominent in trait-oriented research, but it is implicit in any research that seeks to identify enduring individual differences among people, whether or not these differences are viewed as traits. An often overlooked complication of the search for individual differences in children is that developmental differences account for significant variance in almost every measurable behavior. One consequence is that measurements repeated on the same subjects more than a few weeks apart are likely to differ as a function of development, even if the subjects show

stability with respect to their rank ordering within their cohort. A second consequence is that unless all subjects in a sample are at the same developmental level with respect to the behavior in question, individual differences in the behavior may in fact reflect differences in developmental level rather than traitlike characteristics. A third consequence is that covariation among several measures may merely reflect the variance that they all share with development rather than an independent trait.

As an example, impulsivity-reflectivity has been regarded as a dimension of cognitive style that correlates with a variety of behavior in normal and clinical samples. However, it is also known to correlate with CA and MA, indicating that at least some of the variance in measures of impulsivity-reflectivity is shared with indices of development. To determine whether impulsivity-reflectivity represents significant trait variance over and above the variance it shares with cognitive development, Achenbach and Weisz (1975) administered the Stanford-Binet and a test of impulsivity-reflectivity to preschoolers on two occasions 6 months apart. It was found that impulsivity-reflectivity correlated significantly higher with MA than with IQ on each occasion and across occasions. This indicated that cognitive developmental level as measured by MA, rather than deviation from normative age groups as measured by IQ, was the appropriate cognitive measure against which to assess variance in impulsivity-reflectivity.

Although there was a significant correlation between pretest and posttest impulsivity-reflectivity scores, regression of posttest scores on both MA and pretest scores showed that considerably more of the variance in posttest scores could be accounted for by MA than by pretest impulsivity-reflectivity. Furthermore, a significant correlation between impulsivity-reflectivity and hypothesis usage—a behavior assumed central to cognitive style (Kagan & Kogan, 1970)—disappeared when MA was partialled out. It thus appears that general cognitive development may account for much of the variance in impulsivity-reflectivity that has been ascribed to an independent trait. A similar lack of independence from general cog-

nitive development has been reported for field independence (Weisz, O'Neill, & O'Neill, 1975) and for measures of moral development (Taylor & Achenbach, 1975). Before drawing conclusions about variables that correlate with development, it is therefore necessary to demonstrate that they represent reliable variance over and above the variance that they share with general development. Otherwise, we risk a proliferation of "traits" that can be more parsimoniously measured and conceptualized in terms of general indices of development, such as MA and CA.

Avoiding Pathological Biases

It is widely recognized that when people are asked to recall the developmental histories of adults whom the informants know to be disturbed, the reports are likely to be biased in the direction of excessive pathology. Pathological biases of this sort are not restricted to retrospective reports by untrained informants, however. Lacking objective criteria for discriminating normality from pathology, mental health workers may be overly sensitive to signs of pathology. This is especially true where children are concerned, because they may become anxious, constricted, impulsive, or withdrawn when brought to mental health settings, which they view as mysterious, threatening, or punitive. Clinical settings are thus likely to highlight signs of pathology, and the effect of pathological biases on clinical judgments of even the most normal children has been well documented (McCoy, 1976). Research on disturbed children should therefore be grounded firmly on comparisons with normal children matched to clinical samples for conditions of observation as well as for such variables as developmental level, SES, race, and sex.

Correlation, Causation, and Prediction

Causal relationships can rarely be inferred from covariation between variables, unless controlled experimental manipulation of the hypothesized independent variable is followed by corresponding changes in the dependent variable. Experimental designs offer the most powerful means for testing causality, and they

are certainly the most appropriate way to test the effects of variables subject to experimental manipulation, such as treatment conditions. However, true experimental designs are rarely feasible in research on the etiology and developmental course of psychopathology in humans. As a result, much research on psychopathology is correlational. Despite the maxim that "correlation does not imply causation," it is surprising how often statistical covariation is concluded to demonstrate causation without the benefit of experimental manipulation or convergent findings that rule out alternative explanations for the covariation. For example, correlations between child behavior and parent socialization practices have often been interpreted as demonstrating the effects of the socialization practices on children. Yet, as Bell and Harper (1977) have shown, most of these correlations are ambiguous with respect to causation, because they could just as plausibly reflect the effects of children on their parents' behavior, genetic similarities between parents and children, or correlations between the extrafamilial influences that operate on children and on their parents.

Another pitfall of correlational research on children is the tendency to overinterpret correlations in terms of prediction. One source of this tendency is the statistical terminology of multiple regression and discriminant analysis in which the independent variables are referred to as "predictors" of the dependent variable. The summaries of covariation among variables provided by these statistical methods should not be confused with the prediction of behavior across time, however. Unless "predictor" variables are in fact measured earlier than the "outcome" variable and independently of it, they are not predictors in the temporal sense.

A second source of the tendency to overinterpret correlations in terms of prediction is the traditional longitudinal research strategy. A goal shared by research on development and psychopathology is to identify relationships between variables measured at different points in people's lives, and longitudinal studies are the most obvious way to do this. However, significant longitudinal correlations in a particular sample do not necessarily mean

that the earlier variables are predictors of the later variables. Before longitudinal findings are interpreted, it is necessary to determine whether the number of statistically significant relationships substantially exceeds that expected by chance. This is an important step in any research, but longitudinal research offers such inordinate temptations for post hoc analysis that extra care must be taken to control for chance.

Besides controlling for the effects of chance on the distribution of statistically significant relationships, it is important to consider the limitations on generalization of longitudinal findings. Initial selection factors and attrition inevitably undermine the representativeness of longitudinal samples. Moreover, as will be discussed in detail later, peculiarities of particular cohorts and the periods in which they are studied can affect the relationships obtained. Unless they are specifically predicted by a priori hypotheses, covariations between earlier and later measurements in a particular sample are best used to generate hypotheses subject to further test. In rare cases, further tests of the hypotheses may be made with other data from the same samples, or path analysis may be used to choose among competing interpretations (cf. Achenbach, 1978b, chap. 6; Kerlinger & Pedhazur, 1973). More typically, new samples or designs will be needed to replicate and triangulate the relationships obtained in the initial samples. Without replication and further test, significant covariation between earlier and later measurements in a single sample is not a sufficient basis for inferring "prediction," much less causation.

Problems in Measuring Change

At first glance, the measurement of stability and change in behavior appears to be a simple task. However, it is fraught with perils for which there are few simple solutions. I will first consider regression effects and then the problems encountered in using change scores.

Regression effects. Any score can be regarded as a sum of a true score for the variable in question and the errors of measurement that cause the obtained score to deviate

from the true score. Sources of error include momentary fluctuations in the phenomenon being measured, as well as biases and inconsistencies in the measuring procedure. Like most multidetermined phenomena, errors of measurement are assumed to be normally distributed around each of the true scores that would be obtained if there were no errors of measurement. If errors are normally distributed around a true score, extremely large errors will be much rarer than small errors. Therefore, if a large error is made on one occasion, the most probable outcome of a subsequent measurement is a score that is closer to the true score than the first score was. If the first score was extremely high, then the subsequent score will be lower. On the other hand, if the first score was extremely low, the subsequent score will be higher.

To illustrate the way in which regression effects can complicate research, suppose that 100 subjects all have a true score of 50 on variable X. Due to errors of measurement, the actual scores obtained by these subjects on a particular occasion may be normally distributed from 40 to 60, with a mean of 50. Suppose that the same 100 subjects are measured again 6 months later, that their true scores have all remained at 50, and that their obtained scores are again normally distributed from 40 to 60, with a mean of 50. Unless the error of measurement is perfectly reliable from the first to the second occasion, a plot of the relations between each subject's first and second scores will show that subjects who scored lowest on the first occasion now score higher. Conversely, subjects who scored highest on the first occasion now score lower.

Next, suppose that we had performed an experiment to compare the effects of a particular treatment on subjects who initially scored between 40 and 45 on variable X with the effects on those who scored between 55 and 60. Or suppose we could not introduce an experimental manipulation but simply wished to compare the developmental course of subjects who had initially scored low and high on variable X. In either case, we would find the Time 2 scores of both groups to be much closer to the overall mean of 50 than they were at Time 1, merely because the classification of subjects was based on measure-

ment errors that were replaced for those particular subjects by less extreme measurement errors at Time 2. If instead of comparing low and high scorers we had studied the effects of a particular treatment on low scorers, we would have found their Time 2 scores to approximate the mean for the entire sample and might erroneously conclude that the treatment was responsible for the increase in their scores.

Errors of measurement are not the only source of regression effects. Behavioral variables are influenced by so many determinants that even true scores are subject to regression effects from one occasion to another. This is because the lowest true scores on variable X are obtained by subjects who have received the most extreme combination of X-depressing influences. By contrast, the highest true scores on variable X are obtained by subjects who have received the most extreme combinations of X-enhancing influences. Because extreme combinations of influences on X are likely to be due at least in part to chance, the true scores of extreme scorers are likely to regress toward the population mean on subsequent occasions if X is a variable whose true value changes. Thus, children who display many behavior problems on one occasion are likely to display fewer problems on a subsequent occasion due to regression effects. Since most behavioral variables are subject to change—especially variables measured during the course of development—regression in true scores must be expected in addition to regression due to errors of measurement.

Change scores. It would seem logical to compute the difference between an initial score and a later score to measure developmental change over time, as well as change in response to experimental manipulations. However, the size of each change score depends not only on the effects of intervening development and manipulations but also on regression effects. Thus, subjects initially obtaining low scores on a variable are likely to have the largest positive change scores, whereas subjects initially obtaining high scores are likely to have the largest negative change scores.

Even though regression effects may differentially affect change scores of individual high

and low scorers, this would not bias a comparison of groups receiving two different treatments, *provided* that the groups receiving the different treatments had the *same* distribution of initial scores. However, if one group initially scored higher than the other, regression effects could enhance the change scores of one group while depressing the change scores of the other group. Even if change scores are not used, other considerations argue against testing the effects of independent variables with groups differing in initial scores on the dependent variable. Yet, because change scores mask both the initial size of scores and regression effects, they complicate the problem still further.

The most general solution to problems of measuring change is to compare groups that are initially well matched on all relevant variables, including their scores on the dependent variable. Random assignment, stratified randomization, and formation of matched blocks within which treatments are randomly assigned provide the most straightforward ways to obtain comparability. When any form of matching is used, however, care must be taken to avoid biased selection from populations whose distributions differ. For example, parents of lower-SES boys tend to report more delinquent behavior problems than parents of upper-SES boys (Achenbach, 1978a). In order to match lower-SES boys to upper-SES boys on delinquent behavior, it would be necessary to select lower-SES boys whose behavior is less delinquent than the average of their population and/or upper-SES boys whose behavior is more delinquent than the average of their population. However, matching of this sort is fallacious, because, besides being unrepresentative of their SES groups, the boys would be likely to show divergence toward the means of their respective populations on subsequent occasions due to regression effects. Detailed treatment of statistical approaches to the analysis of change is beyond the scope of this article, but advice on compromise solutions to measuring change has been provided by Campbell and Stanley (1963), Cronbach and Furby (1970), Harris (1963), McCall and Appelbaum (1973), and Wilson (1975).

Effects of Age, Cohort, and Time of Assessment

In addition to general problems in measuring change, assessment of behavior that can change with age is complicated by the fact that age differences may be confounded with characteristics peculiar to the cohorts studied and with the effects of cultural conditions prevailing at the time of assessment. Confounding of this sort is most evident when longitudinal conclusions are drawn from cross-sectional data. After considering sources of error in this practice, I will outline various research strategies for separating the effects of age, cohort, and time of assessment. (See Achenbach, 1978b, chaps. 4 and 7, for a more detailed presentation.)

Drawing longitudinal inferences from cross-sectional data. Because longitudinal research is so slow and expensive, it is tempting to use cross-sectional data to infer age changes in behavior. It might seem reasonable to assume that if 6-year-olds in a population behave differently from 16-year-olds in the same population, then the 6-year-olds will change until, when they reach 16, they will resemble the current 16-year-olds. Depending on the behavior in question and the ages spanned, this can be a very risky assumption. Differences in the birth order, SES, cultural experience, health, and schooling of the two cohorts can have powerful effects on their developmental course. Differences in all of these variables and many more can occur even when subjects appear to belong to the same population because they live in the same locality and attend the same schools. The recent decline in the birthrate, for example, means that children who are currently 6 years old will have fewer siblings when they are 16 than do current 16-year-olds. Although the effects of cultural trends are difficult to measure directly, the myriad of uncontrolled variables operating on each cohort makes it risky to infer that cross-sectional age differences reflect the same changes in behavior as occur when a single cohort is studied longitudinally.

One of the clearest contradictions between cross-sectional and longitudinal findings has emerged from the study of IQ test performance at various ages. The norms for the

Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955), for example, were based on cross-sectional samples that showed progressive declines in performance from early to late adulthood. By contrast, longitudinal data on samples of adults retested at intervals of 12 years have shown significant increases in WAIS performance (Kangas & Bradway, 1971). Likewise, a cross-sectional study of the Binet IQs of southern black children showed progressively lower scores from the younger to the older cohorts (Kennedy, Van De Riet, & White, 1963). However, retesting of the same children 5 years later showed that their IQs had remained stable (Kennedy, 1969). In both cases, the declines in IQ implied by the cross-sectional data were due to differences in cohorts other than their ages: In the WAIS samples, the older cohorts had grown up when average educational attainment was less than for the younger cohorts. In the southern black sample, the cohorts were from 5 to 16 years old but included only children who were in elementary school. This means that the youngest cohorts contained children who were bright enough to begin school early, but the older cohorts contained progressively larger proportions of students who had not been promoted beyond elementary school at the customary age.

Jensen's approach. In a widely publicized alternative to the ordinary cross-sectional design, Jensen (1977) has made cross-sectional comparisons between siblings to test the hypothesis of a cumulative deficit in the IQs of southern black children. He has done this by computing the differences between the IQs of black siblings to determine whether older black children consistently have lower IQs than their younger siblings. He then compared these sibling differences in IQ to those of white children in the same schools. Because the younger and older cohorts were from the same families within each racial group, Jensen hoped to control for the possible effects of cohort differences in gene pool, SES, and other factors that might arise at various ages. Consistent with the cumulative deficit hypothesis, Jensen found that younger blacks generally had higher IQs than their older siblings. The sibling differences were less consistent for whites, indicating no general trend

toward cumulative deficit. Jensen concluded from this cross-sectional finding that the IQs of blacks but not whites were declining with age.

Unfortunately, Jensen's alternative leaves so many possible variables uncontrolled that it does not improve much on standard cross-sectional designs. Without passing judgment on the cumulative deficit hypothesis per se, it is instructive to consider weaknesses of the Jensen study that argue against drawing longitudinal conclusions from even this revised version of a cross-sectional design. First, Jensen at one point describes his sample as including "all of the white and black children enrolled in the public schools of a small rural town" (p. 185), but then he says that "the present data . . . include only subjects ranging in age from 6 to 16 years" (p. 187). Since the school system probably enrolled children younger than 6 and older than 16, Jensen's unreported procedure for excluding subjects could have exerted a bias like that apparent in the Kennedy et al. (1963) cross-sectional study. The inclusion of siblings would not entirely eliminate such a bias. There is also no report of alternative schooling opportunities for either whites or blacks that might cause differential attrition from the public schools at different ages. Nor is there any report of how sibling status was ascertained or of how the distribution of half-siblings and stepsiblings might differ between the races. All of these factors could influence the results of cross-sectional sibling comparisons.

The primary comparisons between the races involved analyses by family size such that the mean IQ differences were computed between each child and each of his/her younger siblings in families having 2, 3, 4, and 5 children, respectively. However, very few white families had more than 3 children. In fact, all four comparisons for 5-child families involved only 2-4 white sibling pairs. Besides the obvious statistical fallacies of reporting and testing means with such small samples and with repeated representation of the same families in these samples, the different distribution of family sizes in the black sample and white sample could interact with birth rank in ways that are not controlled by cross-sectional comparison of siblings. For example,

more of the older black children in the sibling comparisons for each family size were likely to be lower in birth rank simply because the black families were larger on the average. Since neither the birth orders nor the ages were reported for children of each race having each family size, these factors cannot be ruled out.

The failure of Jensen's (1977) sibling-based cross-sectional approach to control for so many variables does not by itself argue for specific alternative explanations for the differences between IQs of black children and their younger siblings. However, because Jensen's approach leaves so many variables uncontrolled, the findings do not enable us to choose among a variety of possible explanations. As an example, one possible rival interpretation of Jensen's findings is that recent improvements in living conditions for black families have produced higher IQs in their most recently born children than in their children born before conditions improved. Nothing in Jensen's data can rule out alternative interpretations of this sort, and neither conventional cross-sectional designs nor Jensen's approach are likely to discriminate between age changes and stable differences between age cohorts.

Separating the effects of age, cohort, and time of assessment. Despite the relative advantages of longitudinal over cross-sectional designs for identifying changes with age, longitudinal designs can also produce misleading results if there is selective attrition of subjects or if only one cohort is studied. Research on a single cohort may obscure the fact that what appears to be an age-related change in behavior is either peculiar to that cohort or is manifested by other cohorts at that point in history, regardless of their ages. In a 2-year longitudinal study of adolescents, for example, Nesselroade and Baltes (1974) found that cultural changes from the beginning to the end of the study had more influence on personality than did age changes. The cohorts ranged from 13 to 16 years in initial age and 15 to 18 in final age, but all four cohorts showed declines in superego strength, social-emotional anxiety, and achievement. As a result, the youngest subjects resembled the oldest subjects at the

end of the study rather than resembling the oldest subjects as they *had* been at the beginning of the study. If a single cohort had been studied longitudinally, it might have been erroneously inferred that the changes in personality were a function of age per se, whereas comparison of the cohorts showed that the changes occurred simultaneously in all four cohorts.

Because age, cohort, and cultural-historical effects may be confounded in conventional cross-sectional and longitudinal designs, it is important to separate these variables as explicitly as possible in any studies of development, normal or abnormal (cf. Schaie, 1965). As an aid to separating these variables conceptually, the relevant relationships between them are illustrated in Table 1. Table 1 shows that a cross-sectional study comparing 5-, 7-, and 9-year-olds in 1978 would require children from cohorts born in 1973, 1971, and 1969, respectively. A threat to the internal validity of this purely cross-sectional design is that differences between the 5-, 7-, and 9-year-olds might be attributable to characteristics other than age. A threat to external validity is that any similarities or differences among cohorts might not be generalizable to earlier or later points in time because of cultural-historical changes.

As illustrated in Table 1, a longitudinal study of the 1975 birth cohort at ages 5, 7, and 9 would require assessing the 1975 cohort in 1980, 1982, and 1984. The internal validity of this design is vulnerable to the possibility that differences found from 1980 to 1982 and 1984 might be due to cultural-historical changes rather than to aging per se. The external validity may be limited by peculiarities of the cohort that distinguish it from cohorts born earlier or later.

A third design, the time-lag design, is sometimes used to identify cultural-historical effects. As illustrated in Table 1, historical changes in the behavior of 11-year-olds could be studied by assessing subjects from the 1969, 1971, 1973, and 1975 cohorts when they reach age 11 in 1980, 1982, 1984, and 1986, respectively. However, this design confounds possible differences in the cohorts with differences in year of assessment.

To avoid confounding age, cohort, and time

Table 1
Interrelationships Among Time and Age Variables Involved in Developmental Analyses

Birth cohort	Age			
	5	7	9	11
1969	1974	1976	1978 ^a	1980 ^b
1971	1976	1978 ^a	1980	1982 ^b
1973	1978 ^a	1980	1982	1984 ^b
1975	1980 ^a	1982 ^a	1984 ^a	1986 ^b

Note. Figures in the table are the years in which each birth cohort would be studied at each age listed. (Adapted from Achenbach, 1978b.)

^a Cross-sectional.

^b Time lag.

^c Longitudinal.

of assessment, several designs have been developed to combine aspects of the cross-sectional, longitudinal, and time-lag strategies. The Nesselroade and Baltes (1974) study of adolescent personality combined these strategies in a *longitudinal sequential design*. By assessing several birth year cohorts over the same longitudinal period, it is possible to do cross-sectional comparisons of cohorts at any point in the study, longitudinal comparisons of each cohort, and time-lag comparisons among cohorts as they reach a particular age in successive years, as illustrated in Table 2. It can thus be determined whether changes in behavior are attributable to cultural-historical changes, age changes, or an interaction between the two.

Another design—the *cross-sectional sequential design*—combines cross-sectional and longitudinal strategies by making cross-sectional comparisons of samples from several cohorts at several different points in time. As illustrated in Table 2, the pattern of analyses is similar to that in the longitudinal-sequential design, except that instead of longitudinally following one sample from each cohort, new samples are drawn from each cohort at each time of assessment. Drawing new samples avoids biases affecting longitudinal studies due to attrition, the effects of repeated testing of the same subjects, and initial selection for expected availability. On the other hand, because new samples are drawn for each assessment, changes in individuals over time cannot be traced, and congruence among suc-

Table 2
*Longitudinal Sequential, Cross-sectional
 Sequential, and Time-Lag Sequential Designs*

Birth cohort	Age				
	5	6	7	8	9
Longitudinal sequential ^a					
1971			1978	1979	1980
1972		1978	1979	1980	
1973	1978	1979	1980		
Cross-sectional sequential ^b					
1971			1978	1979	1980
1972		1978	1979	1980	
1973	1978	1979	1980		
Time-lag sequential ^c					
1971			1978	1979	1980
1972			1979	1980	1981
1973			1980	1981	1982

Note. Adapted from Achenbach (1978b).

^a A single sample from each of three cohorts is assessed on three occasions from 1978 to 1980.

^b Three samples are drawn from each of three cohorts on three occasions from 1978 to 1980.

^c Three samples from each of three cohorts are compared on three occasions from 1978 to 1980, 1979 to 1981, and 1980 to 1982, respectively.

cessive samples may be reduced by unrecognized sampling fluctuations and changes in the composition of the cohorts from which the samples are drawn.

In an additional design, known as the *time-lag sequential* design, two or more samples are assessed from each cohort as they reach different ages in different years. As illustrated in Table 2, such a design could be used to study the behavior of the 1971, 1972, and 1973 birth cohorts as they reach the ages of 7, 8, and 9 in 1978–1982, respectively. This design permits separation of effects due to time of assessment from effects due to age of assessment without requiring longitudinal samples. However, like cross-sectional sequential designs, the time-lag sequential design is vulnerable to variations in samples from the same cohort. The chief advantage of this design is to obtain information on the effects of cultural–historical changes with fewer observations than longitudinal sequential or cross-sectional sequential designs would require.

Outstanding Research Needs

There is no doubt that research on child psychopathology is in a rudimentary stage and that it faces many challenges. Perhaps the greatest challenge is to accept the task of forging a science suitable to the problems despite the primitiveness of our knowledge base and the influence of myth, fad, and custom. Other challenges can be summarized in terms of general needs as follows: the need to view children as continually changing with respect to biological, cognitive, social, educational, and emotional development; the need to assess behavior in relation to development that is normal and adaptive within the child's culture; the need to evaluate current behavior from a longitudinal perspective on what has gone before and what is likely to follow; the need to see children's behavior in a family context over which children have little control; the need to develop alternatives to mental health services that require the recipients to assume the role of patient; and the need to abide by ethical constraints without abdicating our responsibilities to seek knowledge that will benefit troubled children. A thorough discussion of the latter issue is beyond the scope of this article, but it is clear that the pendulum swing from inadequate ethical guidelines to minute regulation of everything labeled *research* presents formidable obstacles to the study of child behavior (cf. Achenbach, 1978b, chap. 10). However, lacking validated means for helping troubled children, we must consider not only the ethics of specific research procedures but, to paraphrase Haywood (1977), the ethics of *failing* to do the research needed to improve our ways of helping children. In addition to the general needs facing research on child psychopathology, a number of needs specific to our current stage of development are outlined in the following sections.

Standardized Measures

The handicaps stemming from our lack of well-validated and standardized measures have already been alluded to, but insufficient effort is currently devoted to remedying this situation. Although highly sophisticated in-

umentation requires well-refined and validated theory, it is equally true that theory is unlikely to progress without reliable descriptions of the phenomena in question. The development of standardized descriptive instruments for behavior is a slow, arduous, and unglamorous enterprise that is effectively discouraged by our graduate training programs, reward systems for researchers and teachers, professional journals, and funding agencies. In the face of such obstacles, is it any wonder that our standardized instruments are limited primarily to IQ and achievement tests, which owe their existence to subsidies from the educational system?

If it is acknowledged that research on child psychopathology is indeed in a rudimentary stage, it should be clear that development of standardized procedures for assessing pathologies and competencies and for measuring change in them should command a high priority. The more diverse the sources of data (parents, teachers, peers, clinicians, self-reports) and the greater the focus on important life situations, the more useful such procedures are likely to be. It is only when researchers and clinicians can rely on a common body of baseline measures that we are likely to see significant coordinated advances in knowledge.

Type of Subject \times Type of Treatment Designs

A vast quantity of research has accumulated on the outcome of adult treatment. However, the value of this research has been unduly limited by the heterogeneity of its subject samples, without adequate regard for differences among the subjects. The much smaller quantity of research on child treatment shares the same flaw, despite the fact that when subject characteristics have been analyzed, their effects have been found to account for more variance in outcome than do differences in treatment alone (e.g., Love & Kaswan, 1974; Miller, Barrett, Hampe, & Noble, 1972). Because it should be obvious that no single treatment modality will be better than all other modalities for all children regardless of age, sex, IQ, SES, family constellation, and so forth, we should not have to

repeat the history of adult psychotherapy research before we recognize that outcome studies will be most informative if they take account of subject variables and especially of Subject \times Treatment interactions.

Subject \times Treatment designs are, of course, even more difficult to implement than ordinary controlled studies of treatment outcome, which are themselves formidable. However, because the effects of most interventions are likely to be subtle and multidetermined, it is essential that interventions not be viewed independently of the children with whom they are used. Any outcome study that provides systematic comparison of results for children differing in demographic, diagnostic, and other characteristics will be far more valuable than studies that do not.

Long-term Follow-ups

Although by no means infallible, the most comprehensive approach to determining the significance of potential predisposing factors and manifestations of psychopathology in childhood is through long-term follow-ups. It has become evident that even major adult psychiatric disorders such as schizophrenia may be best understood through longitudinal study of children who are at high risk for the adult disorder (Mednick, Schulsinger, Higgins, & Bell, 1974). Because of the relatively low base rate for severe psychopathology in the general population, long-term study of subjects known to be at risk is more efficient than traditional longitudinal studies of broad samples of children. However, studies of broad samples can also be informative with respect to psychopathology if they are sufficiently focused on the context and course of disorders that a large percentage of children manifest at least temporarily (Thomas, Chess, & Birch, 1968). As mentioned earlier, relationships identified in single cohorts studied longitudinally provide an inadequate basis for definitive inferences, but the complexity of human behavioral development and the rudimentary state of our field make longitudinal data a unique source of hypotheses for further test, even if the effects of age, cohort, and time of assessment cannot be fully separated.

In addition to basic longitudinal research, it is important to build a longitudinal dimension into studies of treatment outcome. Unless adequate follow-up periods are used, it cannot be known how interventions will affect the behavioral, social, emotional, cognitive, and educational development of children. Successful adaptation depends not merely on the removal of troublesome behavior but also on the continued acquisition of progressively more advanced coping strategies. Studies that include follow-ups of even 1 or 2 years often show results quite different from shorter follow-ups of the same children (e.g., Hampe, Noble, Miller, & Barrett, 1973). Arduous though they may be, follow-ups over periods of at least a year should become a routine part of clinical services, as well as of research on these services.

Cumulative Programmatic Research

The volume of publications related to child psychopathology far outweighs its actual impact. Although no field shows a linear progression of knowledge, especially during its formative stages, the literature on child psychopathology reflects two problems that may be remediable. One is the preponderance of one-shot studies whose contribution to knowledge is limited by the uniqueness of their procedures, measures, and subject samples. Since almost no study provides definitive answers by itself, a study is of little value if it cannot be linked to others for purposes of replication, generalization, cross-validation, and triangulation of findings.

The preponderance of one-shot studies is not surprising in view of the obstacles to programmatic research encountered by the graduate students, academicians, and practitioners who have been the chief contributors to our literature. However, unless research can be better coordinated with respect to procedures, measures, and subject samples, and unless each study can be meshed better with what comes before and after it, new studies will add more confusion than knowledge. As stated at the beginning of this article, opportunities are needed for career researchers who can pursue programmatic studies more effectively than people who are also carrying heavy

teaching and/or clinical responsibilities. In addition, graduate students and others who are limited to doing one-shot studies should be encouraged by their advisors and by journal policies to link their work as closely as possible to other research through the use of standardized measures, procedures, and subject samples that can be generalized and replicated. By their sanctions against closely linked series of studies, graduate committees and journal editors may otherwise be perpetuating fragmentation where convergence is needed.

A second impediment to cumulative progress is the tendency for practitioners and researchers alike to identify with a single theoretical viewpoint. This has resulted in what might be called a "horizontal" progression within each school of thought, rather than a "vertical" progression, whereby ideas from various sources are tested and revised or discarded according to the empirical findings they generate. As productive as research within a single paradigm may sometimes be, no single theory or level of analysis is yet so powerful for representing child psychopathology that it deserves exclusive allegiance. In the foreseeable future, cumulative knowledge is more likely to emerge from integration of multiple perspectives than from dogmatic adherence to a single one.

Linking Research to Service Systems

The ultimate aim of research on child psychopathology is to aid children by preventing maladaptive development and by ameliorating it when it does occur. Children receiving treatment through the current mental health system represent only the tip of a very large iceberg that encompasses the educational, legal, medical, and welfare systems. When direct mental health services are available for children at all, they are often sought only after the other systems have failed. These systems are ill equipped to identify and correct maladaptive deviations in behavioral development, but they are representative of the social context in which children develop. It is within this context of mixed motives and messages, competing agendas, maldistribution of resources, and bureaucracy that the fruits of research must ultimately be consumed.

None of the systems is designed to foster research, much less to implement its findings. Even though the mental health system should be the most responsive to research on child psychopathology, it raises barriers of its own. It is clear, for example, that the personal philosophies of mental health practitioners determine the treatments received by clients to a far greater extent than in organic medicine. The concomitantly greater difficulty of changing one's approach to treatment of psychopathology than of trying out a new medication for an organic disorder probably accounts for the resistance of many practitioners to participating in research and to adopting new approaches, no matter what research shows about their customary approach. On the other hand, research related to treatment of children has rarely offered practitioners well-validated alternatives to their customary approaches. Is it therefore surprising that little effective cross-fertilization occurs between research and practice?

To promote cross-fertilization as well as to increase the ecological validity and potential application of their work, researchers must gear their work as closely as possible to the needs of practitioners and must provide practitioners with usable results in a constructive manner. It is only through implementation in practice that researchers can obtain feedback on the effectiveness of their efforts.

References

- Achenbach, T. M. *Developmental psychopathology*. New York: Ronald Press, 1974.
- Achenbach, T. M. The Child Behavior Profile: I. Boys aged 6-11. *Journal of Consulting and Clinical Psychology*, 1978, 46, 478-488. (a)
- Achenbach, T. M. *Research in developmental psychology: Concepts, strategies, methods*. New York: Free Press, 1978. (b)
- Achenbach, T. M., & Edelbrock, C. S. The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin*, in press.
- Achenbach, T. M., & Weisz, J. R. Impulsivity-reflexivity and cognitive development in preschoolers: A longitudinal analysis of developmental and trait variance. *Developmental Psychology*, 1975, 11, 413-414.
- American Psychiatric Association. *Diagnostic and statistical manual of mental disorders* (1st ed.). Washington, D.C.: Author, 1952.
- American Psychiatric Association. *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, D.C.: Author, 1968.
- Back, R., & Dana, R. H. Examiner sex bias and Wechsler Intelligence Scale for Children scores. *Journal of Consulting and Clinical Psychology*, 1977, 45, 500.
- Bell, R. Q., & Harper, L. V. *Child effects on adults*. Hillsdale, N.J.: Erlbaum, 1977.
- Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally, 1963.
- Cronbach, L. J., & Furby, L. How should we measure "change"—Or should we? *Psychological Bulletin*, 1970, 74, 68-80.
- Freeman, M. A reliability study of psychiatric diagnosis in childhood and adolescence. *Journal of Child Psychology and Psychiatry*, 1971, 12, 43-54.
- Gerstein, A. I., Brodzinsky, D. M., & Reiskind, N. Perceptual integration on the Rorschach as an indicator of cognitive capacity: A developmental study of racial differences in a clinic population. *Journal of Consulting and Clinical Psychology*, 1976, 44, 760-765.
- Gesten, E. L. A Health Resources Inventory: The development of a measure of the personal and social competence of primary-grade children. *Journal of Consulting and Clinical Psychology*, 1976, 44, 775-786.
- Group for the Advancement of Psychiatry. *Psychopathological disorders in childhood: Theoretical considerations and a proposed classification*. (Report No. 62). New York: Author, 1966.
- Hampe, E., Noble, H., Miller, L. C., & Barrett, C. L. Phobic children one and two years posttreatment. *Journal of Abnormal Psychology*, 1973, 82, 446-453.
- Harris, C. W. (Ed.). *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.
- Haywood, H. C. The ethics of doing research—And of not doing it. *American Journal of Mental Deficiency*, 1977, 81, 311-317.
- Hobbs, N. *Issues in the classification of children*. San Francisco: Jossey-Bass, 1975.
- Jensen, A. R. Cumulative deficit in IQ of blacks in the rural south. *Developmental Psychology*, 1977, 13, 184-191.
- Kagan, J., & Kogan, N. Individuality and cognitive performance. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology*. (Vol. 1, 3rd ed.). New York: Wiley, 1970.
- Kangas, J., & Bradway, K. Intelligence at middle age: A thirty-eight year follow-up. *Developmental Psychology*, 1971, 5, 333-337.
- Kennedy, W. A. A follow-up normative study of Negro intelligence and achievement. *Monographs of the Society for Research in Child Development*, 1969, 34(2, Serial No. 126).
- Kennedy, W. A., Van De Riet, V., & White, J. C. A normative sample of intelligence and achievement of Negro elementary school children in the southeastern United States. *Monographs of the Society*

- for Research in Child Development, 1963, 28(6, Serial No. 90).
- Kent, R. N., & O'Leary, K. D. A controlled evaluation of behavior modification with conduct problem children. *Journal of Consulting and Clinical Psychology*, 1976, 44, 586-596.
- Kerlinger, F. N., & Pedhazur, E. J. *Multiple regression in behavioral research*. New York: Holt, Rinehart & Winston, 1973.
- Levitt, E. E. Research in psychotherapy with children. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis*. New York: Wiley, 1971.
- Love, L. R., & Kaswan, J. W. *Troubled children: Their families, schools and treatments*. New York: Wiley, 1974.
- Marwit, S. J., & Neumann, G. Black and white children's comprehension of standard and non-standard English passages. *Journal of Educational Psychology*, 1974, 66, 329-332.
- McCall, R. B., & Appelbaum, M. I. Bias in the analysis of repeated measures designs: Some alternative approaches. *Child Development*, 1973, 44, 401-415.
- McCoy, S. A. Clinical judgments of normal childhood behavior. *Journal of Consulting and Clinical Psychology*, 1976, 5, 710-714.
- Mednick, S. A., Schulsinger, F., Higgins, J., & Bell, B. *Genetics, environment and psychopathology*. New York: American Elsevier, 1974.
- Miller, L. C. Q-sort agreement among observers of children. *American Journal of Orthopsychiatry*, 1964, 34, 71-75.
- Miller, L. C., Barrett, C. L., Hampe, E., & Noble, H. Comparison of reciprocal inhibition, psychotherapy, and waiting list control for phobic children. *Journal of Abnormal Psychology*, 1972, 79, 269-279.
- Mischel, W. On the future of personality measurement. *American Psychologist*, 1977, 32, 246-254.
- Nesselroade, J. R., & Baltes, P. B. Adolescent personality development and historical change: 1970-1972. *Monographs of the Society for Research in Child Development*, 1974, 39(1, Serial No. 154).
- Novick, J., Rosenfeld, E., Bloch, D. A., & Dawson, D. Ascertaining deviant behaviors in children. *Journal of Consulting Psychology*, 1966, 30, 230-238.
- Quereshi, M. Intelligence test scores as a function of sex of experimenter and sex of subject. *Journal of Psychology*, 1968, 69, 277-284.
- Rose, S. A., Blank, M., & Spalter, I. Situational specificity of behavior in young children. *Child Development*, 1975, 46, 464-469.
- Schaie, K. W. A general model for the study of developmental problems. *Psychological Bulletin*, 1965, 64, 92-107.
- Shepherd, M., Oppenheim, B., & Mitchell, S. *Childhood behavior and mental health*. New York: Grune & Stratton, 1971.
- Taylor, J., & Achenbach, T. M. Moral and cognitive development in normal and retarded children. *American Journal of Mental Deficiency*, 1975, 80, 43-50.
- Thomas, A., Chess, S., & Birch, H. G. *Temperament and behavior disorders in children*. New York: New York University Press, 1968.
- Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation, 1955.
- Weisz, J. R., O'Neill, P., & O'Neill, P. C. Field dependence-independence on the Children's Embedded Figures Test: Cognitive style or cognitive level? *Developmental Psychology*, 1975, 11, 539-540.
- Wilson, R. S. Analysis of developmental data: Comparison among alternative methods. *Developmental Psychology*, 1975, 11, 676-680.
- Zigler, E. The retarded child as a whole person. In H. E. Adams & W. K. Boardman (Eds.), *Advances in experimental clinical psychology*. New York: Pergamon Press, 1971.

Received September 30, 1977 ■

Sex Roles and Psychological Well-Being: Perspectives on Methodology

Judith Worell
University of Kentucky

The purpose of this article is to delineate some sources of problematic methodology in recent research designed to relate current measures of sex-role orientation to indices of psychological well-being. Practices and procedures in sex-role research are examined in relation to orthogonal scales of sex-role orientation that provide independent measures of masculinity, femininity, and a newer assessment of androgyny. Directions for increased conceptual and methodological clarity include theoretical and psychometric definitions of androgyny, the relationship of sex-role typing to other aspects of interpersonal functioning, and varying procedures in sex-role and gender distinction, population sampling, and construct validation. Issues are raised concerning the generality of sex-role measures and the desirability of direct behavioral validation criteria.

New formulations of psychological health and well-being have renewed an interest in the measurement of masculinity and femininity (Bem, 1974; Berzins, Welling, & Wetter, 1978; Constantinople, 1973; Heilbrun, 1976; Spence, Helmreich, & Stapp, 1975). In contrast to earlier conceptions of sex-typed personality descriptions that relied on a single bipolar dimension, recent approaches to assessment of masculinity and femininity view these as independent, orthogonal dimensions. When treated in this manner, characteristics of masculinity and femininity can be measured in varying amounts in the same individual. Persons are considered to be sex typed to the extent that they endorse a relatively high degree of one set of characteristics in preference to the other. The most conceptually productive outcome of the orthogonal model of sex typing is androgyny, whereby the person endorses relatively equal numbers of masculine and feminine traits (Bem, 1974). The

flurry of new scales to measure androgyny has been followed by a storm of research designed to relate responses on these scales to "something else."

The aim of this article is to examine the methodological characteristics of recent research on psychological sex roles, with the intent of clarifying some current practices and procedures in relation to commonly accepted criteria of research design and analysis. Since earlier reviews have considered at length and in detail the psychometric properties of both traditional and recent sex-role scales (Constantinople, 1973; Kelly & Worell, 1977; Worell, Note 1), issues in scale construction will be minimized here. The focus of the present remarks will be on the implications of current sex-role personality measurement and research for conceptions of psychological health and personal well-being.

The theoretical link between sex typing and adjustment has been altered dramatically since the introduction of the androgyny model. Traditional formulations of sex typing suggest that adoption of the sex roles appropriate to one's male or female gender is developmentally desirable. Deviations from culturally sanctioned sex-role behavior were considered maladaptive and undesirable (Kagan, 1964; Kohlberg, 1966; Mussen, 1969). In contrast,

Appreciation is extended to Suzanne Martorano, to Mary Jo Moeschl, and to two anonymous reviewers for their many helpful suggestions in the preparation of this manuscript.

Requests for reprints should be sent to Judith Worell, Department of Educational Psychology and Counseling, Dickey Hall, University of Kentucky, Lexington, Kentucky 40506.

the androgyny model of sex-role organization suggests that a relative balance of sex-typed characteristics may lead to the most advantageous outcomes. It is important to examine the recent research strategies in relation to the differing assumptions that underlie the selection and interpretation of dependent variables. Androgyny has been compared to sex typing in relation to one or more of the following outcomes: (a) adaptive, flexible, and effective interpersonal behavior; (b) self-esteem or positive self-evaluation; (c) freedom from obvious pathology; and (d) broad life-style coping variables. Each of these adjustment criteria will be discussed in the context of current sex-role research. Since this discussion is not intended as a review of the literature, only selected illustrative examples will be included.

Bem (1974, 1975, 1976) conceives of the androgynous person as adaptive, flexible, and effective in particular interpersonal contexts. Accordingly, the androgynous person can be both *instrumental* (assertive, competent, forceful, independent) and *expressive* (nurturant, warm, supportive, compassionate), depending on the demands of the situation. The outcome of an androgynous orientation is a high degree of alternative options for attaining interpersonal reinforcement in situations requiring culturally sex-typed behaviors. On the Bem Sex-Role Inventory (BSRI), androgyny was originally scored as the relative balance, determined by a negligible *t* ratio, between masculine and feminine sex-typed characteristics (Bem, 1974). In contrast, sex-typed persons scored significantly higher on either their masculine or feminine sex-role traits. Bem considers the sex-typed individual to be more constricted and behaviorally limited in situations in which sex-inappropriate, or cross-sex-typed, behavior is required. Individuals who are highly masculine or feminine sex typed may inhibit or suppress these cross-sex behaviors to avoid self- or other-disapproval.

In a series of construct validation studies using the BSRI, Bem and her associates provided some support for these predictions. Sex-typed individuals did show discomfort and self-defeating choices (monetary loss) in cross-sex situations. In addition, the sex-typed

males showed lowered supportive, playful, and expressive behaviors across several situations, and sex-typed females failed to maintain independence under external pressure to conform (Bem, 1975; Bem & Lenney, 1976; Bem, Martyna, & Watson, 1976). These apparent deficits in the behavioral repertoire of sex-typed, in comparison to androgynous, persons lend support to a flexible-restrictive interpretation. If psychological health and well-being are defined in terms of the availability of behaviors for achieving interpersonal satisfaction, then according to Bem, there may be many instances in which sex-typed persons are clearly at a disadvantage.

A second important dimension of psychological well-being entered into the issues surrounding the measurement of androgyny. Using the Personal Attributes Questionnaire (PAQ) of sex-role endorsement, Spence et al. (1975) found that both male-valued and female-valued scores contributed to a measure of self-esteem. These authors suggested that masculinity and femininity may contribute in an additive way to an individual's positive self-evaluation. Consequently, their definition of androgyny included absolute strength as well as the relative balance of masculinity and femininity scores. The decision to include absolute strengths of each sex-role dimension resulted in a fourfold index, determined by a median split of the combined male and female PAQ scores. Here, the androgynous persons are those who score above the median on both masculinity and femininity. A new category, consisting of those who score below the median on both scales, is termed *undifferentiated*.

The utility of the fourfold scoring system for conceptions of positive well-being was demonstrated in the correlations between the PAQ and a measure of self-esteem. When androgyny was defined in terms of both response strength and balance, androgynous persons were the highest in self-esteem, followed by those who scored high masculine-low feminine, or sex-typed masculine. When Spence et al. rescored these data according to the Bem balance method, androgynous persons scored at only a moderate level of self-esteem. The implications of these findings for positive psychological functioning have led to the adop-

tion of a four-quadrant scoring system by Bem (1976, 1977), as well as by developers of two more recent sex-role inventories (Berkins et al., 1978; Heilbrun, 1976). The scoring issue is not a closed one; some researchers continue to use and defend a balance, rather than an absolute strength, definition of androgyny (Jones, Chernovetz, & Hansson, 1978; Wiggins & Holzmüller, 1978). As currently conceptualized, however, most researchers have preferred to include the absolute strength of masculine and feminine characteristics as an indicator of positive and effective interpersonal functioning. From a response repertoire approach to adaptive situational behavior, the person who endorses more positive self-characteristics appears to have higher self-esteem and should be functional interpersonally in a wide variety of culturally sex-typed situations.

A third indication of effective psychological functioning is the extent to which individuals can remain relatively free from obvious pathology or self-defeating patterns of behavior. Sex-role measures have been variously related to such adjustment indices as anxiety, self-criticism, dependency, helplessness, depression, problem drinking, neurosis, introversion, and requests for personal counseling (Baggio & Neilson, 1976; Deutsch & Gilbert, 1976; Gump, 1972; Heilbrun, 1968; Jones et al., 1978). Since androgynous persons are assumed to be more adaptive, they should have better coping skills and be relatively free of problem behaviors. One concern related to this aspect of adequate psychological functioning is the extent to which individuals provide themselves with negative self-communications and negative self-evaluations (Bandura, 1969; Mahoney, 1974; Meichenbaum, 1977). Although at least two of the presently used sex-role scales (BSRI; PAQ) confine their attributes to positive and valued characteristics, the contribution of negative self-attribution to sex-role categorization has implications for all of the adjustment indices listed above (Kelly & Worell, 1977).

Recent evidence suggests that androgynous males, and some androgynous females, endorse the fewest negative self-statements (Kelly, Caudill, Hathorn, & O'Brien, 1977; Wiggins & Holzmüller, 1978). In contrast to these

positive findings, androgynous males were also found to contribute more than their expected proportion to self-reported problem drinking and social introversion (Jones et al., 1978; Wiggins & Holzmüller, 1978). The research findings in this area of adjustment pathology remain cloudy and inconsistent, especially with regard to androgynous males. Sex-role functioning, as defined by the present measures, may have differential prediction for some psychopathological variables, but no consistent pattern has emerged.

Finally, some investigators have hypothesized that the flexibility presumed to operate in instrumental and expressive domains will be manifested in more effective functioning in a wide variety of cognitive and interpersonal life-style variables. Research efforts in this direction have attempted to relate sex-role types to such variables as creativity, liberal political views, endorsement of feminist views, sexism, marital adjustment, parenting skills, multiple personality factors, and preferred coital position. In some of these studies, the theoretical relationship between effective psychological functioning and androgyny seems quite tenuous. The general theme seems to be that androgynous persons ought to be better at everything, because they are presumably more flexible and adaptive. When androgyny is approached from this viewpoint, it frequently slips into becoming a contest between white hats and black hats.

The discussion of the theoretical underpinnings of current androgyny research becomes important when we take a look at some of the methodological problems that beset the research efforts in this field. In the remainder of this article, I wish to consider the more glaring examples of questionable practices, as well as some of the better ones, in the conceptualization, design, and analysis of research into the nature of psychological sex roles. In doing so, examples will be drawn as prototypes from both published and unpublished research and will be unreferenced. The purpose here will be to organize and discuss these problems in the spirit of constructivism, rather than to point a finger at any particular culprits. The intent of this discussion is to inject a dose of preventative thought and reflection at the early planning stages so that

small problems in design and presentation can be remediated or obviated.

Problems and Practices

The following discussion will generally approximate a theory, design, procedure, and analysis format. Specific topics to be covered are (a) application of theory to design; (b) sex-role and gender distinctions; (c) sampling practices; (d) test and construct validation procedures; and (e) psychometric and statistical considerations. A final section will deal with some unresolved issues.

Theory and Rationale

A substantial portion of manuscripts on sex-role issues fail to explicate a theoretical foundation for the research. This is especially true when a variety of tests or measures are administered. Frequently, little consideration is given to the constructs being measured and their proposed relationship to any sex-role theory. In particular default are studies that take a "one-shot" or a "statistical dragnet" approach. In the one-shot approach, the researcher tends to oversimplify the task of construct validation, implying that any one instance or situation can possibly confirm or negate a psychological concept. For example, suppose a researcher finds that sex-role categories do not predict which women wish to be addressed as Ms., rather than Miss or Mrs. Does this finding negate the utility of the androgyny construct? In a similar mode, the statistical dragnet technique correlates a variety of tests and measures with scores on a sex-role scale. The empirical analysis then capitalizes on a handful of unpredicted results that may disappear in cross-validation. Such atheoretical approaches tend to be devoid of sustained consideration for the particular sex-role-relevant characteristics they are measuring and what they expect to find. The result is frequently a *mélange* of miscellaneous studies and unrelated information that contribute little to progress in current sex-role research.

Although it is reasonable to admit our ignorance in the early stages of exploring androgyny, surely there are some organizing

hypotheses that prompted the research in the first place. Even a replication study might discuss briefly (a) why this study is useful for the procedures it intends to replicate, (b) what it expects to accomplish with this particular sample, and (c) what the implications and/or limitations are in advance of what the results might uncover. What has occurred in the area of sex-role research is a rash of new scales and a bevy of eager researchers who wish to relate androgyny to everything and anything.

There also appear to be a number of confusions and disagreements concerning the theory behind the androgyny research. Bem, for example, limited her conceptions of flexibility to situations that require effective, interpersonal, and culturally sex-typed behavior. Suppose a researcher wishes to extend this notion of flexibility to preferences for differing coital positions. Here, it is important to specify in advance the evidence for any particular position to be sex-role appropriate or inappropriate. Only in this manner can we conclude that sex-typed persons tend to inhibit cross-sex behavior or that androgynous persons can comfortably engage in particular practices. Likewise, if one wishes to correlate sex-role types with acceptance of the women's liberation movement, what is the basis for predicting a favorable response from androgynous individuals? If a man describes himself as warm and supportive, does this necessarily mean that he is willing to assume responsibility for dirty dishes and diapers? Perhaps this is a viable prediction, but again, it requires considerable support in advance. This issue also touches on the distinction between sex-role traits and sex-role behaviors. Within the present context there is no *a priori* reason why sex-role traits cannot be predictive of a broader band of sex-role behaviors than those encompassed by the instrumental and expressive domains. In practice, however, the basis of this type of extended prediction is frequently unclear, making the results uninterpretable.

Cross-cultural and contrast-group studies can be particularly fertile for application of androgyny theory to sex-role research. Since the scales currently in use were constructed and standardized on middle-class American

college students, hypotheses can be tested on how subcultures and differing vocational or interest groups compare. In these contexts, it is the responsibility of the researcher to propose some hypotheses about pertinent differences in cultural norms or sex-role practices and how these might effect the obtained results. Sex-role stereotypes in Formosa or Brazil may indeed differ from U.S. responses, but how do the obtained data articulate with the cultural standards or role demands of each culture? Similarly, there are implicit population hypotheses embedded in an assessment of sex-role responses of homosexual persons that may or may not be logically derived from a sex-role theory. Findings that indicate no differences in sex-role endorsement between specified groups and published college norms may have important implications for the theory or they may be allocated to the "so what" catch-all box.

The inappropriate application of theory to a particular measurement instrument, to a class of behaviors or traits not directly related to the theory, or to miscellaneous contrast populations may result in a premature rejection of hypotheses derived from the theory. If an investigator uses a sex-role measure to predict to a wider band of behaviors than those encompassed by the theory, or to a tangential class of responses, negative findings do not negate the theory. For example, an investigator wishes to test the hypothesis that androgynous persons are more adaptive than sex-typed individuals. A sex-role scale and an anxiety scale are administered; androgynous males turn out to have higher anxiety scores than sex-typed males. Does this finding imply that androgynous males are less, rather than more, adaptive? And does this finding point to a hole in the theory? When using a particular scale, or dependent measure, to assess predictions from a specific androgyny scale, it is extremely important to differentiate the scales (e.g., PAQ, BSRI), the construct (androgyny), and the theory (adaptability). In this case, the negative findings do not necessarily imply a useless theory, a questionable construct, or an invalid sex-role scale. A competing hypothesis might suggest that the higher anxiety of androgynous males reflects their awareness of their unorthodox match

with culturally prescribed standards for masculine behavior. The high obtained anxiety scores for androgynous males may expand our information about this sex-role group, but it does not invalidate previous findings that androgynous males are more willing than sex-typed males to engage in cross-sex behaviors.

How can these problems in relating theory to design be avoided? In the planning and conceptualization stage of sex-role research, the investigator might consider the following: (a) Relate the present research topic to previous findings; (b) state clear experimental hypotheses; (c) relate these hypotheses to some aspect of sex-role theory; (d) state expected outcomes for each hypothesis; and (e) indicate the conditions and limitations under which the hypotheses might support or fail to support the theory. Although these five steps may not eradicate all of the problems in application of theory to design, they may help to prevent unproductive forays and dead-end journeys.

Sex-Role and Gender Distinctions

Judging from the types of designs and statistical analyses that appear in current manuscripts, disagreements exist on how to deal separately with the effects of sex role and gender. Here, gender refers to categorical distinctions between males and females regardless of their behaviors. Sex roles and sex typing will be applied to cultural expectations about the attitudes, beliefs, and behaviors associated with masculinity and femininity. Constantinople's (1973) landmark review presented clear arguments for rejecting gender differences in response as a criterion for item selection or scoring procedures for masculinity-femininity scales. The *raison d'être* of the four sex-role scales referenced here was the measurement of sociocultural sex roles or traits independent of their distribution by gender.¹ Although gender and sociocultural

¹ Two of the four scales under discussion, the PAQ and Heilbrun's Adjective Check List (ACL), have components of gender differentiation built into the structure of the scale. The PAQ contains a separate Masculinity-Femininity scale on which items have been selected by gender discrimination. The

sex roles are not completely independent of each other, they are by no means isomorphic. This distinction has been violated or ignored in at least three major ways.

First, we would like to know whether the functional properties of sex-typed roles or traits are similar for males and females. Some studies have divided subjects by sex type alone, ignoring the Sex Type \times Gender interactions that have appeared in the literature. Some of the obtained interactions between sex types and gender may have distinctive adjustive implications for males and females. Likewise, cross-sex typing appears to function in different ways for the two genders in some studies. The early findings on the functional relationships between gender and sex roles are fragmentary and scattered. Many studies fail to isolate enough cross-sex-typed persons to analyze, but the specific results are suggestive. Further research should clearly include both gender and sex-role considerations.

A problem related to that of the Sex Role \times Gender interaction is the failure to specify the gender distribution of subjects who make judgments or attributions of sex-typed behavior (32 counselors in training were asked to judge . . .). In a similar vein is the failure to analyze for gender even when it is specified and available as an independent variable (36 counselors, 14 males and 22 females, were asked to rate . . .). Surely we are throwing away much valuable data that should be examined rather than swept under the rug. In either instance, the resulting data will not be very meaningful.

The second violation of the Sex Type \times Gender dichotomy is the treatment of data according to gender rather than sex role in assessing or determining sex roles and traits. The effects of sex role and gender should be carefully separated so that we can determine the contributions of each variable. Consider the following examples: To demonstrate that two tasks are sex-role appropriate, the investigator shows that these tasks discriminate the choices of males from those of females. A second example reflects a similar Gender \times Sex

Role contamination. Males and females are assigned to sex-role groups on the basis of a median-split procedure computed with separate male and female frequency distributions. Similarly, sex-role groups are compared on a dependent measure, such as the Minnesota Multiphasic Personality Inventory or introversion-extraversion, the *T* scores for which were computed using separate male and female distributions. In each of these examples, the implications for the resulting data and interpretations may differ. Especially in the latter case, this practice may alter the value of some of the criterion measures and may distort the data in the direction of artificial gender differences. These practices may mask sex-role differences, or Sex Role \times Gender interactions, or they may have no significant effects at all. In each case, the problem can be resolved by examining the data in both ways. Having agreed that sociocultural sex roles and gender are not isomorphic, let us not throw them both into the same pool so that their separate effects become diluted and indistinguishable.

A third confusion concerns the stated purpose of current sex-role scales. The purpose of these instruments is to measure sociocultural sex roles or traits orthogonally and independently of gender. The criterion of relative independence has been the demonstration of a low correlation between the masculinity and femininity scales on each measure. With some exceptions, each scale meets this criterion. Now, an investigator repeats the correlation procedure, supports the scale independence, and concludes that the test is invalid because it fails to discriminate males from females. Or similarly, finding that masculinity and femininity scales on a measure of androgyny do differentiate males from females to some extent, the conclusion is again reached that the sex-role scales are bipolar rather than two-dimensional and orthogonal. In each instance, the purpose of the scales is misunderstood.

Gender and sex-type confusions can be avoided primarily by reversing the procedures above that produced the original contamination. (a) When using sex role as an independent variable, include gender as another variable and analyze for each variable separately.

ACL items were originally selected by differentiation between males who identified with masculine fathers and females who identified with feminine mothers.

In addition, look for possible interactions (providing that the design is appropriate for interactions to appear). (b) Compute median splits for assigning sex-role groups by combining male and female data into a single distribution. (c) Use raw scores on dependent variables in place of *T* scores to avoid criterion problems. (d) Read relevant references, such as Constantinople (1973), Bem (1974, 1977), and Spence et al. (1975), before embarking on any new sex-role research. (e) A writing style addition to this topic is in order. Use a consistent notation to refer to gender (male, female) or to sex roles (masculinity, femininity, androgyny). Frequent changes within an article in notation can leave the reader confused.

Sampling Procedures

Adequate sampling procedures are critical to all areas of psychological research. Certain sampling problems appear particularly relevant to sex-role research because the socio-cultural definitions of sex types and sex-appropriate behavior may vary across populations, generations, and settings. College sophomores participating in experiments for course credit present the fewest sampling challenges; they are available in both genders, in sufficient numbers, and are relatively homogeneous on many relevant characteristics. Restricting all of sex-role research to college students, unfortunately, leaves us with many unanswered questions about the generality of results and the applicability to contrast populations. Those researchers who brave the real world to measure other kinds of people face many more hazards in meeting the requirements for an adequate sample. Therefore, it is particularly incumbent on researchers who leave the college classroom to make certain that their samples meet at least four requirements: Make sure that (a) sample characteristics and selection procedures are recorded and clearly described, (b) samples are sufficiently large and (c) representative to allow generalization to new groups, and (d) the criteria for selection of contrast populations are clearly defined. In the absence of any one of these requirements, the experimental findings may be difficult or impossible to inter-

pret, and the results may have little applicability to other populations. Examples of each of these may clarify the position.

Insufficient descriptions of sampling procedures and population characteristics guarantee that a research design cannot be appropriately replicated. To relate a set of findings to a larger body of sex-role research, it is important for the reader to be informed of the exact *N*, age or age range, gender, and whatever other demographic characteristics of the participants are deemed relevant. For sex-role research, information on career, marital, and educational status might be important. Readers also wish to know how the subjects were enlisted and what the conditions were under which some subjects dropped out or were eliminated from the sample (mortality). Some recent examples of inadequate descriptions include such statements as "20 housewives returning to school were administered . . ." or "30 graduate students at University X in Calcutta were approached and asked. . . ." These vague sample descriptions do not help us to accumulate a body of useful data from which to generalize either to similar or to contrast populations. In cross-cultural research, in which the impact of national characteristics is examined, the appropriate match between cross-national samples becomes critical.

A similar caution is appropriate when examining sex typing in clinical groups such as male alcoholics or female depressives. Unless the demographic characteristics of the clinical population are carefully described and matched with an appropriate control group, sex-role differences may be ascribed to the dynamics of the clinical syndrome that are probably a function of the restricted or atypical sample. For example, a researcher who finds a disproportionate number of feminine-typed males among alcoholics in a Veterans Administration hospital may draw conclusions about sex typing and substance abuse that are sample related rather than syndrome specific.

Inadequate or unrepresentative samples present two additional problems in drawing generalizations from the obtained sex-role data. Although statistical requirements may frequently be met by small sampling procedures,

application to larger populations is questionable. For example, suppose the researcher wishes to examine the effects of counselor gender and sex type on counselor and client behavior during assertion training. If only one counselor from each gender and sex-role category is used, the outcome is particularly vulnerable to unreliable conclusions.

Sampling problems also appear in the selection of control or comparison groups. Research designs that cut across age ranges, such as in developmental comparisons, are singularly sensitive to differences in cohorts, socioeconomic class, parental education, cultural milieu, and achievement or ability factors. Unrepresentative samples are easy to find. Take any set of high school juniors and compare them on a sex-role measure to groups of college and graduate students. Developmental hypotheses such as age or generation effects cannot be adequately tested between these sets of subjects unless each relevant external factor is accounted for, controlled, or covaried statistically. Socioeconomic factors may likewise differentiate an experimental from a normal control group. If a researcher wishes to examine psychological sex roles of girls with unplanned pregnancies who are now living in a home for unwed mothers, what is an appropriate comparison population?

Criteria for the selection of contrast or control groups are frequently unclear or gratuitous. The appearance of new ways to measure sex typing has encouraged the application of these scales to diverse populations that vary along unknown dimensions. Certainly, when contrast groups are used to demonstrate the discriminant validity of a sex-role measure, considerable support should be marshaled for assuming that these groups differ in sex-type orientation. Are all women who choose not to work feminine sex typed and all policemen masculine sex typed? If a validation study finds no difference in sex-role endorsement between these two groups, where does the fault lie? Is the test not valid, or are the groups unrepresentative of each sex type?

The foregoing comments are not meant to underestimate the difficulties in obtaining appropriate comparison groups for various populations. Simple suggestions about how to avoid sampling problems will necessarily leave

many issues untouched. However, consistent attention at least to the four criteria outlined above would be a welcome start.

Construct and Test Validation Procedures

Four major criteria for test and construct validity appear in the research literature: (a) self-report; (b) interpersonal perception and attribution; (c) task and situational performance; and (d) contrast populations. Since contrast populations were considered in the previous section, they will not be covered again. For purposes of brevity, both construct and test validation procedures will be considered together. However, it is important to point out that scale and construct validation do not always involve converging procedures (Fiske, 1971; Wiggins, 1973). Establishing the validity of the BSRI should be a less exhaustive procedure than considering the theoretical and conceptual domains encompassed by the construct of androgyny. For example, the inter-scale correlations among the four sex-role scales discussed above range from .61 to .81 for masculinity and from .51 to .73 for femininity (Kelly, Furman, & Young, in press). These correlations shrink by 52%–58% when the same correlations are computed for dichotomized scores (above and below the median). Although each scale is designed to measure some aspect of androgyny, it is clear that they overlap only partially and may be assessing somewhat differing conceptual and content domains. Consequently, the predictive validity of each scale needs to be assessed individually and independently of the scale validity of the remaining measures. Nevertheless, the following discussion will cover criterion variables as they apply both to the validity of specific tests and to the construct validity of androgyny, recognizing that some comments may apply differentially to each.

Self-report studies are by far the most popular method of demonstrating test or construct validity for androgyny. Self-report is obviously quick and economical, and it has face validity. Self-report methods base their strength on the belief that persons are their own best observers (Mischel, 1968; Mischel, Note 2). Problems arise with self-descriptive measures in androgyny research when these

scales are (a) insufficiently described; (b) marginally related to scales previously used to measure a similar construct; (c) dubiously related to the construct of androgyny or to the domain of sex typing; (d) unpublished, unvalidated, and possibly unreliable; or (e) in violation of the separation of gender and sex typing. When any of these shortcomings appear, questions can be raised about the significance of the data reported using self-report measures. Scales with unknown psychometric properties are particularly vulnerable when negative results are obtained. If the investigator finds no relationship between androgyny scores on the PAQ and "a scale especially contrived to measure openness to experience," is androgyny theory at fault, is the PAQ not predictive of "openness," or do we have an invalid criterion? A further problem in restricting the measurement of dependent variables to self-report is the introduction of substantial "method variance" (Campbell & Fiske, 1959). For example, to what extent does the obtained positive correlation between masculine and androgynous sex types and measures of self-esteem reflect the tendency of these people to endorse a high degree of many positive traits?

Despite their shortcomings under certain circumstances, self-report measures of dependent variables are not likely to disappear from use. Therefore, investigators who use self-report scales can refine their approach by considering one or all of the following: (a) Give preference to scales that have been used previously in similar research; (b) describe the scales used in sufficient detail to inform the reader of relevant psychometric characteristics of the scale and its previous use in similar contexts; (c) provide a theory or rationale for use of these particular scales, including their relationship to other scales that have been used to measure the same construct (such as feminism or need for achievement); and (d) develop directional hypotheses about the probable relationship between measures of sex typing and the variables presumably tapped by other self-report measures. Progress in sex-role theory depends on a body of reliable and comparable research. New data are meaningless unless they can be fitted into a matrix of accumulated knowledge. When di-

vergencies do occur, it is the task of the researcher to support the use of particular measures and to indicate their relationship to existing methods.

Interpersonal ratings and attributions represent a second area of validation research. Studies using this approach generally fall into one of two categories: (a) Subjects rate selected others on sex-typed characteristics to obtain stereotypes or (b) clinical analogues are constructed in which subjects evaluate written, visual, or auditory material about standard stimulus persons who appear in several sex-typed roles. The most popular variation of stereotype ratings has been to attempt to replicate the study by Broverman, Broverman, Clarkson, Rosenkrantz, and Vogel (1970), mainly to demonstrate negative outcomes. A common difficulty in replication of sex-role stereotype studies is neglect to repeat exactly the procedures and instructions used by the earlier studies. For example, in a single study, researchers change the target person from adult to child, the raters from clinicians to child-care workers, the instructions from "healthy" to "ideal," the scoring procedures from seven to two rating categories, and they shorten the number of items in the scale. Although this type of study may have utility for some purposes, it is clearly not a proper replication of Broverman et al.

A more appropriate application of replication procedures might be systematically to vary each factor that could contribute to the total variance in stereotyped attributions. In this manner, perhaps the researcher could substantiate a hypothesis that previous results are a function of instructions, item format, scoring criteria, or a sampled population. For example, should a researcher seriously wish to examine stereotyping in differing populations, it would be useful to obtain more than one contrast group. Thus, to show that teachers do or do not hold these stereotypes is probably not exciting enough to make a worm turn these days. However, it might be interesting to determine stereotyping differences among groups that vary along a particular dimension such as education, employment, administrative power, and so forth, especially when the chosen dimension has implications for institutional or administrative change.

Clinical analogues in which standard stimulus persons are given diagnoses, labels, or prescriptions for behavior change have crept into sex-role research. Since analysis of analogue research is beyond the scope of this article, comments will be limited to sex-role implications. A typical study runs like this: Subjects, usually undergraduates, are given a standard personality description and are asked to provide a diagnosis, recommend treatment, or make predictions about marital adjustment. Sex roles enter into the design when stimulus persons vary on gender and sex-role-appropriate or sex-role-inappropriate behaviors. Suppose the outcome of this design shows that male and female stimulus persons who use sex-role-reversed behaviors are judged as more seriously ill, more in need of treatment, and less likely to be happy in marriage. What are the problems here? First, generalization about current diagnostic and treatment practices from college sophomores to practicing clinicians is questionable. Do judgments on a two-paragraph written description replicate referral behavior in clinical situations in which multiple sources of information are available and when economics and family relationships are involved? Second, sex-role implications for each stimulus person may be unequal. When John runs sobbing into the bedroom after a marital quarrel, is this equal in social value to Mary who tuned up the family car in preference to going out for lunch with her friends? Problems in task equivalence appear in performance studies as well as in attribution or analogue studies. In either case, it is necessary to equate, by means of observer agreement, the sex-role relevance of the task or situation and its social value (degree of positive or negative valence). In the absence of this information, we cannot conclude that sex-role inversions lead to dire social consequences.

Two related areas of attribution research involving sex roles center on achievement variables: fear of success and achievement motivation attributions. The issues related to these topics are far too complex to consider here, but a number of recent reviews are available (see Frieze, 1975; Stein & Bailey, 1973; Tresemer, 1976; Zuckerman & Wheeler, 1975). It should be noted, however, that many

of the same problems occur in all of these attribution areas: task variables, expectancy for and the value of success, and external validity.

Task performance and behavior samples represent the most direct external validation procedures in sex-role research. Aside from what people tell us they do or prefer, or what others judge them to be, how do sex-typed or androgynous persons actually behave in sample situations? Although there are many problems involved in making behavioral predictions from self-reported characteristics (Fiske, 1971; Mischel, 1968; Spence & Helmreich, 1978; Mischel, Note 2; Spence, Note 3), behavior samples are essential to sex-role validation research. Androgynous orientations should predict greater adaptive or effective behavior across situations designed to elicit behaviors in the instrumental or expressive domains (Bem, 1974, 1975). Persons who describe themselves in few sex-typed terms might be expected to show behavioral deficits equally across sex-role correlated situations (Kelly, O'Brien, Phillips, Hosford, & Kingsinger, Note 4). Individuals who score high on either instrumental or expressive domains might be expected to show competence on sex-typed tasks relevant to expressive and instrumental domains and to exhibit behavioral inhibition or deficits on similar sex-reversed tasks.

Clearly, the conceptual and experimental challenge here is to construct tasks and situations that match the theoretical predictions. The two most common problems appear to be (a) criterion or construct validity: Is the selected situation congruent with the construct? Does the task represent a reasonable and logical extension of the construct domain to be examined (e.g., conformity, competitiveness, nurturance, risk taking)? and (b) sex-role validity: Does the task represent a logical sample of sex-role behavior within the domains assessed by the particular sex-typing measure? Issues in the general criterion validity of tasks are common to all research (Fiske, 1971) and will be omitted here. Sex-role validity of tasks has been considered only recently in the literature (Stein & Bailey, 1973) and is still problematic in current research efforts. One conceptual challenge con-

task selection is to delineate the probable domains of observable behaviors encompassed by the instrumental and expressive dimensions. To my knowledge, none of the current sex-role scales was designed to measure all of the traits that discriminate males from females or to differentiate all sex-typed cultural traits. When predictions to external behavioral situations extend beyond the conceptual network of the assessment instrument, low or negligible results can be expected.

The experimental challenge involves, among other things, careful attention to task selection for the behavioral validation of sex-role measures. Considerations should include adequate pretesting of tasks for sex-role endorsement; the value of the tasks for the participants; and the expectancy for success on the tasks, equated across sex-role or gender groups. In many studies, no information is provided on how the tasks were chosen. The resulting tasks may be gender typed, sex role typed, or only tangentially related to either of these variables. For example, when examining sex typing on the choice of luck or skill games, an electronic ball game may not be the most neutral task. It probably elicits both value and expectancy differences among groups. Any conclusions that are drawn about tendencies of sex-typed or androgynous persons to prefer skill or luck tasks are immediately biased by the nature of the activity.

If current sex-role measures are to receive a fair trial, pretest procedures should clearly differentiate tasks according to a sociocultural and sex-role definition (what is more desirable, typical, appropriate, etc.), rather than by gender discrimination (what differentiates males from females). Although gender selection may be used to demonstrate gender typing once the tasks are chosen, it seems appropriate to restrict the final task selection to a cultural judgment criterion. Comparisons between these two procedures may show that the basis on which any set of tasks is chosen has no appreciable effect on the data. However, this comparison is seldom attempted. Meanwhile, many researchers continue to contaminate gender and sex roles in design and task selection, thereby confusing and masking the contributions of each.

A summary of current practices in the vali-

dation of sex-role measures reveals a number of design problems that may render the results uninterpretable. In all three areas reviewed here—self-report, attributions, and behavior samples—procedural difficulties appear that are potentially avoidable.

Psychometric and Statistical Considerations

Discrepant approaches to data analysis are not limited to sex-role research. However, certain procedures are sufficiently problematic and common that they merit a brief commentary. Two topics will be considered; scoring for androgyny and statistical treatment of data.

Scoring procedures that are problematic fall into three categories: (a) failure to report how scales were scored or to report mean scale scores (A study reports the number of sex-typed and androgynous subjects but does not explain how these categories were obtained.); (b) idiosyncratic scoring systems that do not demonstrate the effects of these changes on the data (Researcher figures up a new way to calculate androgyny and reports only resulting data, without comparison to previous scoring criteria.); (c) misuse of androgyny scores indicating some misunderstanding of their meaning. (Study reports that one group was twice as androgynous as the other, based on the absolute differences between the androgyny scores.) Any one of these scoring practices prevents the reader from relating the present research to previous findings. New scoring procedures may be extremely useful, but their utility should be demonstrated by means of data-impact comparisons between the new and the traditional methods.

Statistical procedures deserve more than a passing nod. However, since many sex-role studies attempt to demonstrate group differences, a word of caution is justified. By far the most common annoyance is data analysis that treats each dependent variable separately, using individual *t* tests, chi-squares, one-way analyses of variance, or columns of single correlations. Aside from the opportunity to capitalize on chance variations, many dependent variables in personality research tend to be correlated. Single tests of significance cannot correct for the common variance in a group

of correlated measures. Many research studies would benefit from the use of multivariate procedures, and readers would benefit by encountering fewer dramatic results that fail to hold up under cross-validation.

Conclusions

The wealth of current research designed to evaluate the correlates and consequences of androgyny has been beset with multiple methodological problems. Many of the design and procedural difficulties that appear are common to other areas of virgin research. However, the current orthogonal conceptualization of sociocultural sex roles and their measurement by specific scales that meet this assumption introduce a number of methodological issues that are specific to androgyny investigations. As a function of Bem's unveiling of androgyny as a model of effective psychological functioning, validation studies have centered on the adjustment advantages of a relatively balanced, as compared to a sex-typed, sex-role endorsement. The resulting research efforts have produced some very creative solutions to the challenges of construct validation, as well as some repetitive and inconclusive approaches.

At the risk of oversimplification, one major design and conceptual problem centers around a diffusion and confusion between gender (male, female) and sex-role endorsement (masculine, feminine, androgynous). The distinction between these two variables is frequently ignored in the development of hypotheses, tasks, and tests. Even though Bem and her associates (Bem, 1975, 1976) have been careful to make this distinction clear, many others have not. The position was taken here that all components of androgyny research should differentiate cultural sex roles from gender typing. This consideration implies that theory, measurement of both independent and dependent variables, as well as reporting procedures, would benefit by respecting the sex role/gender distinction. Only in this manner can we tease out the relative contributions of sex-role advantages or drawbacks to males and females separately.

The limited research data accumulated so far suggest that sex roles emphasizing either

instrumental, expressive, or androgynous orientations may have differing implications for the psychological well-being of males and females in American culture (Bem & Lenney, 1976; Berzins et al., 1978; Jones et al., 1978; Kelly et al., 1977; Wiggins & Holzmuller, 1978). Although these group differences tend to favor the presence of moderately high masculinity scores (either alone or in a combined androgyny score), the adjustment advantages of masculinity are by no means universal and are tempered by the situation, the task, the measures used, and the gender of the experimenter. In a recent series of studies across a diversity of situations, Jones et al. (1978) found a number of Gender \times Sex Role interactions, with androgynous males showing more adjustive problems and both androgynous and masculine-typed females appearing more adaptive. The implication of these diverse findings is clearly to direct attention to refinements in research design, a specification of limiting conditions under which particular results can be expected to occur, and a continued careful distinction at all levels of research between sex-role and gender effects.

A second major issue relates to the psychometric definition of androgyny. At the present time, considerable disagreement exists concerning the appropriate method for translating raw scores on current sex-role scales into a predictive metric that is both statistically sound and psychologically meaningful. Two major scoring procedures are currently in use: Bem's original t ratio, which emphasizes an intraindividual subtractive balance definition of androgyny (Bem, 1974), and a median-split procedure, which includes the absolute numbers of masculinity and femininity item endorsements within a specified sample (Spence et al., 1975). Comparison of data using either of these two procedures suggests that under some conditions the two scoring methods produce equivalent results, and for other conditions the additive method yields new data (Bem, 1977; Berzins et al., 1978; Jones et al., 1978; Wiggins & Holzmuller, 1978). Some researchers have combined the two scoring approaches to produce a new metric based on both the relative and absolute scores (Babl, in press; Heilbrun & Pitman, Note 5). Still others have suggested the use

multiple regression analysis to take into account the full range of scores on both masculinity and femininity scales (Kelly et al., in press; Wakefield, Sasek, Friedman, & Bowles, 1976). With linear data, a multiple regression design might overcome some of the measurement limitations of current scoring methodologies. However, there do not appear to be any published data that examine the empirical implications of regression analysis for androgyny research. It is clear that some gains and losses accrue to each scoring procedure, and the issue is by no means closed. Hopefully, accumulated data will suggest which method yields a useful scoring procedure that will enable research from different laboratories to be compared. In the interim, it is suggested here that researchers specify carefully their scoring assumptions and procedures.

Regardless of the scoring scheme adopted, differences in measuring instruments will also contribute to variations in the androgyny validation theme. Although all four of the present scales assessing orthogonal sex roles and traits share a common variance, they are by no means interchangeable. It has been pointed out elsewhere (Kelly & Worell, 1977; Spence & Helmreich, 1978) that these scales all vary in assumptions about androgyny, item selection (origins of item pool, instructions to item raters, criteria for item inclusions, etc.), item content and format, and instructions to respondents (yes-no ratings on bipolar or unipolar choices). In addition, the masculinity and femininity scales show varying correlations with corresponding scales on each of the other tests, although the associations tend to be moderate to high. Moreover, when the same respondents fill out all four sex-role inventories, a large majority (61% when corrections are made for chance agreements) are categorized discrepantly by any pair of scales (Kelly et al., in press).

What these differences in scale characteristics and concordance imply for androgyny research is that generalizations concerning adjustment and psychological well-being should be confined to the particular instrument used to assess sex roles and the particular scoring methods adopted. None of the present scales is sufficiently calibrated so that predictions

from one scale can be transferred to another instrument. A recent study using matched predictions from the BSRI and the PRF ANDRO scales showed very little overlap on a set of dependent variables (Hicks, 1977).

A final issue concerns the extent to which the characteristics measured by any of these current sex-role scales reflect unitary traits or dispositions that are predictive of a wide range of behaviors, attitudes, and life-style choices. The limited data accumulated thus far strongly suggest that endorsement of one's typical degree of instrumental and expressive characteristics is not necessarily predictive of any or all sex-role and gender-correlated behavior. It should come as no surprise to anyone familiar with personality assessment that behavioral prediction from self-described personality traits is constrained by what one is willing and able to do in any particular situation. Self-attributions will thus interact situationally for each person with the value of an activity, expectancies for success, fear of failure, desires to please self and others, and real or perceived behavioral skill. Specific role-determined behaviors may not coexist with trait descriptions, because they vary on one or all of these limiting dimensions.

Of particular interest to androgyny traits is the question of the relative contribution of social competency to performance on sex-typed tasks. The behavioral activities selected to test androgyny hypotheses have included tasks that involve trait-related skills (such as positive and negative assertiveness), as well as role behaviors that require no skill but that may be differentially acceptable or aversive (e.g., diapering a baby, playing with a kitten). Future research should attempt to tease out the extent to which interpersonal skills and the positive and aversive stimulus values of criterion tasks interact with measures of sex-typed orientations. Although it is apparent that orthogonal sex-role measures are not indiscriminately predictive of varying indices of psychological well-being, direct behavioral tests must continue to be important criteria of what people are willing and able to do. As Mischel has aptly pointed out (Mischel, Note 2), situation-free personality variables hold little interest or utility for any psychologist.

Reference Notes

1. Worell, J. *Issues in the measurement of sex roles*. Paper presented at the meeting of American Psychological Association, Chicago, August 1975.
2. Mischel, W. *On the future of personality measurement*. Paper presented at the future of personality measurement symposium, American Psychological Association, Chicago, August 1975.
3. Spence, J. T. *Traits, roles and the concept of androgyny*. Paper presented at the conference on Perspectives in the Psychology of Women, Michigan State University, East Lansing, May 1977.
4. Kelly, J. A., O'Brien, C. G., Phillips, J., Hosford, R. L., & Kinsinger, E. *Interpersonal consequences of sex-typing and androgyny: A behavioral analysis*. Paper presented at the meeting of Southeastern Psychological Association, Hollywood, Florida, May 1977.
5. Heilbrun, A. B., Jr., & Pitman, D. *Testing some basic assumptions about psychological androgyny*. Manuscript submitted for publication, 1978.

References

- Babl, J. D. Compensatory masculine responding as a function of sex role. *Journal of Consulting and Clinical Psychology*, in press.
- Baggio, M. K., & Neilson, E. C. Anxiety correlates of sex-role identity. *Journal of Clinical Psychology*, 1976, 32, 619-623.
- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Bem, S. L. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1974, 47, 155-162.
- Bem, S. L. Sex-role adaptability: One consequence of psychological androgyny. *Journal of Personality and Social Psychology*, 1975, 33, 634-643.
- Bem, S. L. Probing the promise of androgyny. In A. G. Kaplan & J. P. Bean (Eds.), *Beyond sex-role stereotypes: Readings toward a psychology of androgyny*. Boston: Little, Brown, 1976.
- Bem, S. L. On the utility of alternative procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1977, 45, 196-205.
- Bem, S. L., & Lenney, E. Sex-typing and the avoidance of cross-sex behavior. *Journal of Personality and Social Psychology*, 1976, 33, 48-54.
- Bem, S. L., Martyna, W., & Watson, C. Sex-typing and androgyny: Further explorations of the expressive domain. *Journal of Personality and Social Psychology*, 1976, 34, 1016-1023.
- Berzins, J. I., Welling, M. A., & Wetter, R. E. A new measure of psychological androgyny based on the Personality Research Form. *Journal of Consulting and Clinical Psychology*, 1978, 46, 126-138.
- Broverman, I. K., Broverman, D. M., Clarkson, F. E., Rosenkrantz, P. S., & Vogel, S. R. Sex-role stereotypes and clinical judgements of mental health. *Journal of Consulting and Clinical Psychology*, 1970, 34, 1-7.
- Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Constantinople, A. Masculinity-femininity: An exception to a famous dictum? *Psychological Bulletin*, 1973, 80, 389-407.
- Deutsch, C. J., & Gilbert, L. A. Sex-role stereotypes: Effect on perceptions of self and others and on personal adjustment. *Journal of Counseling Psychology*, 1976, 23, 373-379.
- Fiske, D. W. *Measuring the concepts of personality*. Chicago: Aldine, 1971.
- Frieze, I. H. Women's expectations for and causal attributions of success and failure. In M. T. S. Mednick, S. S. Tangri, & L. W. Hoffman (Eds.), *Women and achievement: Social and motivational analyses*. New York: Wiley, 1975.
- Gump, J. P. Sex-role attitudes and psychological well-being. *Journal of Social Issues*, 1972, 28, 79-92.
- Heilbrun, A. B., Jr. Sex-role, instrumental-expressive behavior, and psychopathology in females. *Journal of Abnormal Psychology*, 1968, 73, 131-136.
- Heilbrun, A. B., Jr. Measurement of masculine and feminine sex-role identities as independent dimensions. *Journal of Consulting and Clinical Psychology*, 1976, 44, 183-190.
- Hicks, A. *Psychological sex role as related to vocational choice and certain personality variables*. Unpublished doctoral dissertation, University of Kentucky, 1977.
- Jones, W. H., Chernovetz, M. E. O'C., & Hansson, R. O. The enigma of androgyny: Differential implications for males and females? *Journal of Consulting and Clinical Psychology*, 1978, 46, 298-313.
- Kagan, J. Acquisition and significance of sex-typing and sex-role identity. In M. L. Hoffman & L. W. Hoffman (Eds.), *Review of child development research*. New York: Russell Sage Foundation, 1964.
- Kelly, J. A., Caudill, M. S., Hathorn, S., & O'Brien, C. G. Socially undesirable sex-correlated characteristics: Implications for androgyny and adjustment. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1185-1186.
- Kelly, J. A., Furman, W., & Young, V. Problems associated with the typological measurement of sex roles and androgyny. *Journal of Consulting and Clinical Psychology*, in press.
- Kelly, J. A., & Worell, J. New formulations of sex roles and androgyny: A critical review. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1101-1115.
- Kohlberg, L. A cognitive-developmental analysis of children's sex-role concepts and attitudes. In E. E. Maccoby (Ed.), *The development of sex differences*. Palo Alto, Calif.: Stanford University Press, 1966.
- Mahoney, M. J. *Cognition and behavior modification*. Cambridge, Mass.: Ballinger, 1974.
- Meichenbaum, D. *Cognitive behavior modification: An integrative approach*. New York: Plenum Press, 1977.

- Ischel, W. *Personality and assessment*. New York: Wiley, 1968.
- Kussen, P. H. Early sex-role development. In D. A. Goslin (Ed.), *Handbook of socialization theory and research*. Chicago: Rand-McNally, 1969.
- Loch, J. T., & Helmreich, R. *The psychological dimensions of masculinity and femininity: Their correlates and antecedents*. Austin: University of Texas Press, 1978.
- Loch, J. T., Helmreich, R., & Stapp, J. Ratings of self and peers on sex-role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 1975, 32, 29-39.
- Loch, A. H., & Bailey, M. M. The socialization of achievement orientation in females. *Psychological Bulletin*, 1973, 80, 345-366.
- Tresemmer, D. The cumulative record of research on "fear of success." *Sex Roles*, 1976, 2, 217-236.
- Wakefield, J. A., Jr., Sasek, J., Friedman, A. F., & Bowden, J. D. Androgyny and other measures of masculinity-femininity. *Journal of Consulting and Clinical Psychology*, 1976, 44, 766-770.
- Wiggins, J. S. *Personality and prediction: Principles of personality assessment*. Reading, Mass.: Addison-Wesley, 1973.
- Wiggins, J. S., & Holzmuller, A. Psychological androgyny and interpersonal behavior. *Journal of Consulting and Clinical Psychology*, 1978, 46, 40-52.
- Zuckerman, M., & Wheeler, L. To dispel fantasy-based measures of fear of success. *Psychological Bulletin*, 1975, 82, 932-946.

Received October 19, 1977 ■

Some Problems in Community Program Evaluation Research

Emory L. Cowen
University of Rochester

Realities of the community context militate against good program evaluation research. Many limiting factors in such research stem from a clash in values between those who must deliver and those who must evaluate community services. Detailed consideration is given to several clusters of difficulties that plague community program evaluation studies, including (a) sources of data bias, (b) issues of design, (c) problems in the choice and use of criteria, and (d) problems of experimental control. Although community program evaluation studies can surely be improved, it is unlikely that the purity of antiseptic, laboratory research will ever be attained. Ultimate conclusions about the effectiveness of community service programs may thus have to come about slowly and cumulatively, based on convergent findings from many individual less-than-ideal outcome studies.

The author is well qualified to write about research errors in community mental health because he has, personally, committed virtually every one of them. Had the editor told authors to start their article with a pithy folk wisdom that captured its essence, mine would have been: "Do as I say, not as I do!" This article is intended as a straightforward consideration of several hazards that plague community program evaluation research. It is *not* designed to lecture or to pontificate. If the words come through as "holier than thou," it will reflect a serious communication failure.

Starting this article as I did was dictated by other than modesty—or even masochism. It was to suggest that some problems of community research are so intrinsic to the nature of the beast that they are very difficult to surmount. As stated elsewhere (Cowen, Lorion, & Dorr, 1974), the choice for investigators in this field is often between doing far less than ideal research or no research at all.

With that as preamble, a first practical task

is to sharpen the article's focus—somewhat easier said than done. Terms such as community research or research in community psychology or mental health are broad and amorphous (Cowen, 1973). In an earlier article, I suggested that although clinical psychology, community mental health, and community psychology share common prime concerns with people's adjustment, adaptation, security, happiness, self-concept, that is, their well-being, the nature and timing of their defining practices differ radically (Cowen, 1977). Thus, clinical psychology—as well as psychiatry and social work—has traditionally used repair strategies such as psychotherapy addressed to already evident, crystallized problems. Community mental health's (CMH) roots lie in keenly felt dissatisfactions with the effectiveness of classic mental health repair systems. CMH does not abandon the casualty repair orientation; rather, it directs its efforts toward the perceived insufficiencies of past traditional approaches. The thrusts of the CMH movement are to identify problems earlier, in more natural settings (e.g., schools), and to use more flexible, hopefully more realistic approaches sometimes carried out by non-traditional help agents. The real importance of community in CMH is that it harbors settings and contexts that make it easier to do

My thanks to two anonymous, but gracious and constructive, referees who reviewed the initial draft. Their wisdom and perceptiveness is reflected in the current incarnation.

Requests for reprints should be sent to Emory L. Cowen, Department of Psychology, University of Rochester, Rochester, New York 14627.

these things. Community psychology, sharing much the same goals as traditional or CMH approaches, departs sharply from them in its strategies. It is mass oriented rather than individual oriented, and it seeks to build health from the start rather than to repair.

The preceding account is grossly oversimplified. It extracts distilled essences of pure models that are blurred and muddled, in nature, at many overlapping points. Moreover, the approaches are depicted unidimensionally when, in fact, each consists of an agglomerate of strategies, techniques and, ultimately, research terrains. Bloom (1977) documented that complexity for CMH, identifying 10 important ways in which such approaches differ from past traditional mental health practices. Of special interest to this discussion is Bloom's opening definition of CMH as "all activities undertaken in the community in the name of mental health" (1977, p. 49). In that vein, he cites as a first feature, which distinguishes CMH from traditional clinical activities, the fact that the former are based on *practice in the community*. Several of Bloom's later key discriminanda (e.g., emphases on early service delivery, indirect services, use of non-traditional manpower) are natural derivatives of the community locus of CMH programs.

Important as the preceding structural emphases are, they do not yet begin to identify CMH's substantive complexities. The latter can be illustrated simply by noting several of the field's active current areas of programming and research: needs assessment surveys; alcoholism and drug-abuse programs; mental health consultation by, and for, diverse groups; varied types of crisis intervention programs; selection, training, and performance of nontraditional help agents; informal help-giving processes, natural caregivers, and community support networks; early detection and intervention; and alternative service delivery modes for inner-city and rural folk.

A brief account such as this could not encompass the myriad of research problems of all the foregoing areas even if the author had (which he does not) the expertise to do so. That limitation, however, is more merciful than tragic. Though CMH evaluation research has clear defining qualities, it is not, thank goodness, a world unto itself with a totally

unique technology, methodology, and modus operandi. Its basic problems are related to those of other major areas of outcome research (e.g., evaluation of psychotherapy or educational programs).

The full array of research problems includes those of designing, conducting, and interpreting studies. The greatest commonality in such problems between outcome research in CMH and other areas—and thus the topic *least* pursued in this article—is error in interpreting findings. Thus, overinterpreting chance findings (e.g., the "meaning" of five F s out of 100 significant at $p \leq .05$); confusing statistical significance with meaningfulness in interpreting correlation coefficients (e.g., $r = .07$, significant at $p \leq .05$, because $N = 1,000$); capitalizing on chance to develop a regression equation that is not cross-validated; or overreacting to a significant chi-square based on small expected cell frequencies, unmodified by Yates's correction, are common errors of data interpretation in many substantive areas—not just CMH.

A key thesis of this article is that certain errors in designing and conducting CMH program evaluation studies flow naturally from the special hazards of *doing* research in the community. The latter include (a) the low priority that program evaluation research may have in an agency's hierarchy of values; (b) the researcher may be—or be *seen* as—a foreign body to the system; (c) the threat that evaluation research poses for a program's funding or personnel; (d) difficulties in gaining entry into community systems; (e) the complex research demands that longitudinal programs impose; (f) the vulnerability of community programs to change after they start; (g) the involvements of community review bodies; and (h) growing concerns about human rights and the invasion of privacy. Specific research problems considered later in this article can stem, directly or indirectly, from such qualities of community contexts. Accordingly, many gut problems of community research occur less because investigators do not "know any better" and more because reality keeps them from doing any better.

One key "common denominator" obstacle to sound community evaluation studies is that the researcher's and agency's goals frequently

work at cross-purposes. Schools are for teaching; courts are for meting out justice; community mental health center (CMHC) clinics are for working with patients, and hospital wards are to care for the sick. Community agencies share the mandate of bringing needed services to people. Their first concern must be with the extent and quality of those services. Their prime goal is to insure delivery of optimal services. The evaluator's first allegiance, by contrast, is to sound design, methodology, and instrumentation, which are so important to the defensibility of a study's conclusions. These perspectives can, and do, clash. The real rub comes when program personnel see the design of a prospective study as encroaching on, or restricting, services, and the researcher sees service pressures as a factor that will corrupt the study's integrity. To complicate the matter further, some service personnel see program evaluations as personal evaluations and, therefore, as threatening. Others, not sympathetic to research in the first place, or who feel harassed by heavy job pressures, resent the extra burdens that an evaluation study places on them.

Whether or not a just God in heaven might judge such perceptions to be warranted is not the critical issue. Because they are real to beholders, they result in very real behaviors such as lack of cooperation, anger, passive aggression, delayed and/or careless completion of forms in ways that limit their validity, and interpretability. There are few guaranteed solutions for such problems. Involving program personnel in planning studies, taking time to explain the purposes and significance of the research to them, using maximally parsimonious measures that insofar as possible, are relevant to the respondents' job turf, and feeding back findings in usable forms are steps that can strengthen program evaluation studies. But even so, the differing needs and values of program workers and evaluators remain as background noise that often works against sound research.

That less-than-idyllic backdrop frames consideration of specific problems in CMH program evaluation research in the sections to follow. Although not all of these problems fit neatly into pigeonholes, broad categories such as data and design bias, choice of criteria, and

the misuse of controls, clearly are chronic offenders.

Data Bias

In the past several decades, psychological research in general has become more sensitive to, and sophisticated about, sources of data bias. Observer judgments are highly subject to stylistic inputs such as social desirability, responding, "yeasaying" or "naysaying," excessive use or avoidance of extreme ratings, and halo effects, any of which can be confounded with substantive variables that an instrument purports to assess. Although ways of instrument construction and usage have been developed that minimize stylistic confounds (or, at least, identify them after the fact), the corrosive effects of those variables still mar community program evaluation studies. This is so because such studies often depend heavily on the judgments of service recipients and providers as prime data sources.

Few would disagree with the assertion that a client's view of how he/she has done in a program is *one* relevant way to evaluate the program's effectiveness. But there are many reasons why it is misleading to use such data as the *only* way of evaluating a program (Bloom, 1972). For brevity's sake, the issue can be put in concrete, caricatured form. When a program ends, the male client is asked, in 20 different guises: "So, how'd you do?" His response, also in 20 ways, is "Terrific!" Problem No. 1: Did he respond that way because that's how he feels or because he senses that that's what the experimenter wants to hear? Problem No. 2: If he truly *does* feel better, is it due to the program or because he just struck oil or won the Irish Sweepstakes? Problem No. 3: If indeed he does *feel* better because of the program, has his *behavior* changed in a parallel way?

Service providers are also relevant—indeed, important—data sources in program evaluation, because they are uniquely familiar with clients and their everyday behaviors. But they too are sources of potentially serious data bias. They can bias a study, obviously, by not completing essential forms, by completing them carelessly, or by submitting them too late. More subtly, their responses to questions

about client behavior and how it has changed can be shaped by their stake in, and cathexes to, a program, that is, whether they believe that they are *really* being asked to evaluate the client's behavior or their own effectiveness; whether they see the program as theirs or others' and, if the latter, whether they like or dislike those people.

The dilemma is clear. On the one hand, people who staff a program and/or receive its services have observations and information that are highly pertinent to its evaluation. But, for whatever reasons—many extraneous to the program's content and thrust—they do not always respond in those terms. The interpretability and ultimate contribution of rater judgments may be increased by broadening the observational report bases. There is more danger of bias, pro or con, with only one versus several judges or one versus several perspectives (e.g., self, therapist, job, home). If observers with different stakes and perspectives agree about change, it is more likely that real change has taken place. Even more useful, if feasible, is including behavioral anchor points in the overall evaluation net. Correspondences between bona fide behavior change and the judgments of human observers increase one's confidence in the latter as a converging source of evidence in evaluating program effects.

Design Problems

Campbell (1969) has written a sophisticated treatise on program evaluation designs for community research—a topic somewhat beyond the scope of this article. The present discussion focuses on two specific design issues that have created special problems for community research—follow-up and systematic versus representative design.

The purpose of follow-up is to insure that effects observed when a program ends accurately and stably mirror the program's impact. Follow-up data thus solidify generalizations about program effects over time. Such information is important for planning future programs. Immediate postprogram findings can be misleading in several respects. Thus, what seems to be improvement can dissipate over time, because it was not (a) real in the

first place, (b) solid enough to permit the individual to meet life's demands after the program ended, or (c) supported by the post-program environmental context.

Without follow-up, we can also underestimate program effects. Significant experimental-control differences may *not* be found when a program ends, but they may show up 6 months or a year later. Illustratively, a preventive program is developed for children experiencing current life crises who have not yet shown major signs of maladjustment. At the end of the program experimental subjects look much the same as comparable crisis (non-program) control subjects. When evaluated a year later, however, experimentals are found to have maintained sound adjustment, but nonprogram controls have deteriorated behaviorally and educationally. In such a situation one might infer that the program had important innoculative value but that it was too soon for that effect to be detected when the program ended. Without follow-up the community program evaluator is vulnerable to incorrect conclusions and generalizations. He/she may thus either perpetuate a shaky program or dismiss an effective one prematurely.

Brunswik (1947) wrote an informative essay on systematic and representative design in psychological experiments, later applied to research in clinical psychology (Hammond, 1954). Brunswik's main argument was that in order to generalize beyond a study's specific circumstances, one need not only have an adequate subject *N*, which most experiments do, but also experimental conditions that represent, statistically, the universe of circumstances to which the experimenter hopes to generalize.

A simple example would be a comparative study of the effects of male versus female examiners on operant conditioning rates of male and female subjects. Assume that such a study was done with one male and one female research assistant, each of whom ran 50 male and 50 female subjects, for a total of 200 subjects. Although, on the surface, that seems a "reasonable" *N*, the question is, reasonable for what? How useful is it to know that two particular experimenters—one of whom happened to be male and the other female—got different conditioning results either overall or

differentially by subjects' sex? Does such a finding mean that being a male or a female experimenter was *the* critical variable underlying the observed performance differences rather than, let's say, differences in their warmth, verbal styles, cues emitted, or degree of comfort with subjects? Most studies seek to generalize beyond their own literal conditions. For the hypothetical study cited, to generalize about the effects of male and female examiners would require representative sampling along the dimension of experimenter sex. Thus, in effect, the *N* for the study was not 200; it was 1 in each experimenter's group.

Community program evaluation studies are especially vulnerable to problems of systematic versus representative design. Such research often seeks to generalize about large community units (e.g., schools, CMHCs). The researcher, however, may have access to *one* or, at most, a very limited number of such units. Thus a study is conducted to learn how low-income subjects use mental health services within a CMHC and how effective those services are. The study, done at a large urban CMHC, involves a consecutive sample of 500 low-income subjects who sought services for whatever reasons, during the past *X* number of months. The study's main findings are that 54% of the subjects did *not* return after the initial visit and that short-term goal-oriented therapy was found (by whatever criteria) to be the most effective of four treatment conditions studied. Although such information may be very helpful in pinpointing the practices and strengths of the particular setting, a frequent error is to generalize the findings to CMHCs, or CMHCs in large urban settings. The problem in so doing is that the center in question has its own special defining qualities and practices (e.g., inviting/uninviting physical layout; poor/good reception and/or initial interviewing practices; and committed, dedicated short-term therapists), any of which can produce the specific findings obtained. To generalize about large CMHCs requires representative sampling on *that* dimension.

The problem comes up in many guises. An investigator might wish to compare the effectiveness of clinical case versus process consultation. Such a study is well designed in terms of the number and types of groups (e.g.,

teachers, lawyers, clergymen) with whom the approaches were used and the numbers of contacts with each group. However, if there is only one consultant per approach, generalization with respect to the study's main question is drastically restricted. Because there are so many ways (e.g., experience, comfort, and confidence with an approach; personal warmth; verbal facility) in which the consultants could have differed, besides the ostensible variable under study (i.e., *type of consulting approach*), conclusions about the relative effectiveness of the approaches could not be made without representative sampling on the consultant dimension.

Another example would be a comparative study of the attitudes and job satisfactions of mental health professionals and paraprofessionals. One convenient (often the most convenient) way of doing such a study is to recruit fairly sizable *N*s for both groups from a single, large facility that employs, let's say, 20 professionals and 40 paraprofessionals. Assume that the study is done that way and that clear group differences in job satisfaction are found. The investigator concludes that there are basic cross-group (mental health professionals vs. paraprofessionals) differences in job satisfaction. But, again, because the study was done in a single center, the findings are more likely to reflect that setting's particulars (e.g., hours or conditions of employment, salary levels, promotion policies, job security, how positions at various levels are perceived and valued) than *generalized* cross-group differences on the variable in question.

Generalization of research findings depends on representativeness of design on *all* pertinent dimensions. This cannot ordinarily be achieved in one variable systematic design. If a community program evaluation study seeks to reach conclusions that transcend a particular setting, it must adequately sample the situations and variables that are central to its generalization focus, as well as the usual adequate sampling of subjects.

Criterion Problems

The present discussion assumes that classic psychometric problems (e.g., reliability and validity) are well understood at least intel-

actually, and that researchers as well as editors do what they can to keep the faith with respect to them. Hence, this section focuses more particularly on criterion problems that are predisposed by the special nature and pressures of community research. Two such groups of problems are considered: (a) the extent to which criterion measures are appropriate to a study's purposes and (b) pressures on the researcher to use less than optimal criterion measures.

Are the criteria appropriate to the study's purposes? In considering how CMH approaches differ from traditional ones, one frequently noted characteristic is that they involve indirect rather than direct services (Bloom, 1977). Thus, we do consultation with public health nurses, pediatricians, teachers, and clergymen because these groups have extensive everyday contacts with distressed individuals. The rationale for consultation is that upgrading consultees' knowledge and skills helps them to be more effective in dealing with the personal problems that their clientele often bring to them.

Psychologically oriented educational programs for those who are about to become parents, or for the parents of newborn or young children, have a similar rationale. Hopefully, strengthening the knowledge and/or attitude bases of program participants can lead to more facilitating, health-producing child-rearing attitudes and practices. Thus, consultation and mental health education epitomize the structural pattern of indirect service; that is, they both are directed toward groups that have systematic contacts with and are hence in a position to help others. But the programs' *ultimate* concerns are with the beneficial effects that intermediaries have on target people.

Mental health consultation and education seek to enhance the knowledge, feelings, and attitudes of the groups that they touch. It is often assumed that if those variables change positively, constructive change in the behavior of people with whom the intermediary interacts will ensue. Because indirect service programs have direct contacts only with intermediaries, and because first-line changes are easier to get at than once-removed ones, measures of change in consultees or parents are

often used as the *only* basis for evaluating their effectiveness. Even if criterion measures are well selected at that level and convincing evidence of change is found (i.e., consultees or parents are shown to have *enjoyed* a program, *learned* a lot from it, and to have developed more favorable mental health *attitudes*), it cannot be assumed that those changes lead to more effective helping, or growth-supporting, practices.

Kelly (1971) addresses this issue:

The payoff from a consultation program is not only an alteration of the feeling states, belief systems, and aspirations of the consultee, but should also reflect a change in a person's relationships with those significant others who directly participate in his life setting. Therefore evaluation studies should not measure change in attitudes of consultees, nor analyze samples of the interactions between consultant and consultee, nor note changes in the consultee's self-concept, for such attempts at evaluation are not congruent with a conception of consultation as a preventive intervention. . . . If . . . consultation is effective in initiating a change process, then indices of effectiveness should be defined not only by changes in consultee performance, such as the classroom teacher, but by cumulative and successive changes in the behavior of significant others, for example, students in the classroom. . . . When considering research designs to document the effects of consultation . . . it is essential to provide for assessment of the radiating effects of the intervention. . . . An intervention such as consultation can be preventive only if the consultee produces change in significant others. (pp. 114-115)

The point to emphasize is that although there is nothing wrong per se with using waystation criteria such as changes in the knowledge and attitudes of consultees or parents in indirect service programs, such criteria alone are insufficient. So-called instrumental changes, if found, do *not* guarantee that positive behavior changes will follow in the ultimate target group. Without assessing the latter directly, there is the danger that all concerned parties will have had a pleasant, seemingly productive experience, that fails to help the program's ultimate targets.

Although the preceding is a widespread problem, it is not universal. Behavioral consultation (Heller & Monahan, 1977), for example, is often evaluated *only* in terms of specific behavior changes observed in ultimate targets. Moreover, there are examples of (non-behavioral) indirect service programs in par-

ent education (e.g., Hereford, 1963; Glide-well, Gildea, & Kaufman, 1973), in which positive program effects (behavioral and adaptive) were indeed shown for the intended *ultimate* targets—children. Research designs that include measures of changes both in direct recipients and ultimate targets offer an added dimension richness in that linkages between the two levels of change can be explored.

Pressures affecting the choice of criterion measures. Community background contextual factors, such as those discussed earlier, that interfere in general with program evaluation hit especially hard when it comes to selecting and using research criteria. If the basic climate is not conducive to evaluation, resistances to assessment procedures can, and do, develop. Such procedures can all too readily be seen as time-consuming, disruptive, and personally intrusive.

Time-consuming is often defined phenomenologically rather than objectively. The author has had the experience of finding respondents more receptive to a 1-page format that required 5 minutes to complete than to a similar 10-page format that required only 3 minutes to complete. That aside, the key practical concern is that if key responders see a measure as too time-consuming, for whatever reason, it can effectively rule out that measure as a criterion. Take the following example: The main objective of a day-care program is to "resocialize" patients along dimensions such as (a) initiative, (b) self-help behaviors, (c) interaction with peers, (d) interaction with program personnel, (e) outside recreational activities, and (f) outside job activities. The researcher thus believes that judgments by knowledgeable program personnel about changes in specific behaviors, in each of the above subareas, would be among the criteria of choice for the study. Although he/she develops an appropriate instrument to assess 10 specific behaviors per category that requires only 20 minutes per subject to complete, respondents decide they cannot, or will not, give that amount of time to the task. A frequent "compromise" solution in such circumstances is to use "quick and dirty" but vaguer and more abstract global ratings such as "interactions with peers." Although these may be useful, they also lose the richness of

the phenomena under the study, pull for generalized attitudinal responses to the program and its personnel, and narrow the base for evaluating specific program effects or identifying particular program strengths and weaknesses as a guide to its modification and improvement.

The reality bugaboo of the potential disruptiveness of evaluation procedures is another source of noise that materially restricts the researcher's choice of outcome criteria. One might, for example, envision several other data-gathering strategies for the hypothetical study just cited such as (a) direct behavioral observations of subjects to assess variables such as peer interaction, autonomy, initiative, and reactions to program personnel; and (b) tests that provide data from which inferences about such variables could be made. Both approaches present hazards. Apart from the intrinsic complexities involved in developing reliable, valid frameworks for observing and recording behaviors, such procedures are often seen as intrusive and are therefore resisted. "Outsiders" must be introduced to the doings of an ongoing program, a process that program personnel or participants may see as disruptive, threatening, uncomfortable, or just plain not wanted. Similarly, removing subjects from ongoing program activities (e.g., schoolwork) for evaluations, especially time-consuming ones, also elicits resistance—if not to stop the procedure entirely, then to pare it to the bone.

Sensitivity about potential invasion of privacy is another factor that restricts the use of certain criteria in evaluating community programs. A given program may seek to improve people's sexual adjustment; another may be aimed at solidifying disrupted parent-child relationships. But to probe directly in these sensitive (albeit face-valid) areas may be so threatening that the use of theoretically ideal criteria is blocked before the fact.

Zax and Klein (1960) argued that behavioral criteria are often among the best to use in evaluating the effectiveness of mental health interventions. Because behavior, and its disruption, often defines and is at the nerve center of an individual's problem, and because it tends to be objective and palpable, using behavioral indices of change, is both face valid and commonsensical. Unfortunately, the inac-

ability of such data, plus the fact that it may be costly or time-consuming to obtain, has caused it to be underused in evaluating the effectiveness of community programs.

The preceding constraints on the use of criteria put dire pressures on the community program evaluator to compromise. Compromise sometimes means using indirect measures; instruments that are out of phase with the program's goals; measures of unknown or dubious reliability and validity; and vague, global criteria that are difficult to relate to the study's focal variables. Though such problems are far from unique to community program evaluations, they are pronounced in that field. This is an especially tough blow, since variables of prime concern in CMH program evaluation research (e.g., health, pathology, adjustment, coping) are difficult enough to measure even when we have a "clean shot" at them (Bloom, 1977). Stated another way, program evaluation research is limited by the state of the art in assessment. Illustratively, the researcher may be asked to evaluate a program designed to strengthen the self-concepts of preschool children. If, however, a psychometrically sound measure of self-concept for children of that age is not available, the functional choice is between using an unsatisfactory measure of self-concept or a psychometrically more sound instrument that comes as close as possible to assessing that variable.

Special criterion problems come up in outcome studies that cut across groups or settings. Thus, in a study designed to compare the effects of a specific intervention on the attitudes, intellectual performance, and/or adjustment of middle- and low-income groups, it is important that the criterion measures be equally appropriate for the two groups being compared. If not, extraneous factors such as lack of item clarity or inappropriateness of test items for either group might easily be confounded with differential outcomes for the method being evaluated.

Much the same problem can occur in evaluation studies that cut across structurally comparable settings, which, however, use different assessment procedures. Cowen et al. (1974) have addressed this question in the context of a study designed to evaluate the ef-

fectiveness of a multidistrict school mental health program. Because the program was school based, an estimate of the child's current school performance seemed to be one reasonable criterion for evaluating its effectiveness. But going across school districts made it virtually impossible to obtain such a measure. Formerly, the classic "A, B, C, D, E" report card was an answer to the researcher's prayers precisely because it offered a nearly universal metric for evaluating current academic performance. But "them simplistic days" seem to be gone forever. The ancient grading system has been supplanted by a near infinity of variants—single, double and triple checks; red, blue, and green stars; or lions, tigers, and giraffes. Even more perplexing for the aspiring quantifier are the extensive free-prose reports used by many school districts these days to evaluate children. Such reports have been known to cover 20 or more pages and to be as much, or more, oriented to unfamiliar turf, such as identity problems, socialization skills, and self-concept as to the erstwhile, inviolable three Rs. However laudable efforts to find better, less competitive, less accusatory ways to assess a child's school performance are, they make it difficult, if not impossible, for researchers to use report card grades as a performance estimate in cross-district comparisons of program effects. Similar problems are encountered with educational achievement tests. School districts have become more and more individualized about which tests they use, for whom, and when such tests are given. Thus, in designs that cut across school districts, the experimenter may need to combine apples and bananas if he/she hopes to use performance and achievement data.

Such reality factors will continue to restrict the researcher's use of criterion measures. In addition, the intrinsic complexity of many community evaluation studies poses challenges in selecting criteria. Complexity means several things—*obvious* ones such as (a) the multicomponent nature of many community programs and (b) the fact that they seek to affect multiple functions—sometimes differently for different people—and *less than obvious* ones, such as the fact that they usually take place in specific contexts of cost, morale,

attitudes, and expectancies. Though it helps a great deal to know that a program "works," ultimately it is important to disaggregate component effects, separating active from inert ingredients and identifying differential program effects for participants with different characteristics. Because such information is critical for improving programs, criteria should be selected with those realities in mind.

Because programs have multivariate, differential, and changing outcomes, multiple outcome criteria, including behavioral ones wherever possible, should be used to evaluate them. Doing so not only accurately reflects a program's true complexity, but it also reduces the risk of putting all one's eggs in a single delicate criterion basket. Greater use should also be made of unobtrusive and/or unintended outcome measures; they are relatively easy to collect and potentially informative. Thus, an investigator might weep a bit less after finding that children in a school mental health program improved at $p < .09$, rather than $p < .05$, on a problem behavior inventory, were it also established that the principal had 30 fewer disciplinary referrals during the active program period. And, finally, in evaluation studies that require human judgments (by program personnel or target persons), pragmatic decisions to use brief, objective, and easy to understand and handle measures that are relevant to respondents' main interests and spheres of involvement can help the cause immeasurably.

Problems of control. Problems of control, like those of criteria, are basic to most areas of psychological research. Their uniqueness, if any, to community program evaluation research derives from the realities of the community context. Control, in evaluation studies, seeks to insure that changes in behavior and/or performance are due to the effects of the intervention, rather than to potentially confounding variables that can produce similar findings. Being cast as a control is basically a dull, unenviable fate. It commits the "victim" to all the "dirty work" of research (e.g., time loss, intrusiveness, disruptiveness) without direct "pay off"—characteristically defined as needed services. Small wonder that relatively few places jump through hoops to be controls. Since many settings (schools, hospitals, clin-

ics, courts) are unwilling to serve as controls under any circumstance, the theoretical pool of control groups is limited before the researcher ever gets to it.

Moreover, locating "any old" control group is hardly enough to assure a good outcome study. Ideally, experimentals and controls should be matched on a host of variables which, left unattended, could confound the findings. Although the exact nature of these variables depends on the program's nature and purposes as well as characteristics of the subjects, they often include age, sex, intelligence, race, socioeconomic status, and the nature and extent of preprogram maladjustment. Both because many settings eschew the control role and because experimentals and controls must be matched on many control variables, the experimenter may have few degrees of freedom in searching for appropriate community control groups. At the very least, it can be a vexing, time-consuming challenge.

Not surprisingly, then, many program evaluation studies are done without control groups. Such studies depend on within-group pre-post comparisons, which, though sometimes helpful, entail certain risks. For one thing, they do not control for spontaneous or natural change over time; for another, they rely heavily on human judgments of change—by subjects themselves or by program personnel—they are particularly susceptible to distortions such as response bias (Hawthorne, and reverse Hawthorne effects) (Zdep & Irvine, 1970). Still another danger of studies that lack controls is the tendency of initially extreme test scores to regress to the mean on readministration. Such natural regression on the surface, "looks like" improvement and can be confused with it. Finally, base rates for some behavioral criteria (e.g., delinquency rates and employment) that are appropriate in evaluating the effectiveness of certain programs change rapidly over short time periods (e.g., ages 12–14 for delinquency or 16–19 for employment). To use such criteria, without controls or anchoring base-rate data, could seriously distort a study's interpretation (Freeman & Sherwood, 1969).

Another community reality that undermines optimal control is the fact that experimental programs *must* often start at a certain time

and run for X period of time if they are to be evaluated at all. School-based programs, for example, are bounded by the beginning and end of the school year. The program must start when it must start, even if an adequate control group is not available. Following all due effort an approximately satisfactory control group may be located several months later. But, by then, children manifest different patterns of class adjustment problems (either because they are better known or due to normal seasonal variations), and class demographic structures have changed from what they were 2 months earlier. It is difficult to know, when experimentals and controls are assessed at different times of the year, whether scores on key criterion measures such as the preceding ones mean the same thing for the two groups.

Given the realities of the community research scene, many adaptations to, and "solutions" of, the control dilemma have been tried—some voluntarily, some otherwise. The own-control procedure, in which a group goes through an inert preprogram wait period, is designed to bypass the thorny problems of using matched controls. A variant of that approach is for a group to serve simultaneously as a matched control for an experimental group and as an own control for a specified time period, with the understanding that it will later participate in the regular program procedure. Although both of those variants may be useful, they may still present problems, for example, delay in providing services to individuals who need them or the dangers of systematically selecting as time controls subjects with less pressing service needs.

In some situations the most direct route to control is to subdivide a prospective pool of target subjects, within a given setting, into matched experimental and control subgroups. This approach is attractive both because it is convenient and because it may be easier to find a good overall match among within-setting subjects who share similar backgrounds, sociodemographic status, and histories. But there are also reasons why it may not work. Thus, if the program involves needed services, personnel from the setting may argue—sometimes vocally and insistently—that those with greatest need must have first call on ser-

vices. Withholding a promising service from someone who needs it badly to satisfy the niceties of an abstract research design simply will not fly. Indeed, if pushed too hard, it can sound the program's demise—either before it starts or through later noncooperation or active resistance. It takes no genius to imagine the public relations confusion that can ensue. The obvious point is that community research has a real, vital ecological surround—more so than almost any other area of psychological research. That surround must be taken into serious account at all stages.

Another factor that limits the usefulness of within-setting control is intrasetting communication about a program. If, for example, the program intervention involves the use of cotage parents to teach verbal mediational techniques of self-control to residential delinquent adolescents, there is the danger that training procedures and program practices will spill over from experimentals to controls in a given institutional setting. Many community intervention programs involve indirect services such as consultation. If teachers are targets of a consultation program, it is unrealistic to expect that they will apply newly learned skills only to experimental children in their class and *not* to controls, or that they will not discuss useful new discoveries with other (non-experimental) teachers in the building.

Control, as many investigators have learned, can be an elusive phenomenon; that is, "now you see it, now you don't." Seemingly pretty initial matches evaporate in the face of hazards beyond the experimenter's control. Thus an experimenter, who must start a program by a certain date, matches experimental and control groups on all major sociodemographic measures but not on preprogram adjustment or performance status. Because the latter data take longer to collect and score, the experimenter makes the perfectly reasonable assumption that random assignment of subjects will yield approximately matched groups on those dimensions. But it does not! Or, after careful matching is completed, attrition while the program is underway destroys the match. There are many reasons why such attrition occurs. People move, raters judges who furnished predata change jobs; administrative decisions are made to shift individuals out of

a program; needed data cannot be collected or prove to be invalid. And so it goes! If everyone who has been burned by such a problem submitted a brief description of his/her special headache, the resulting compilation of unanticipated, and "undeserved," false turns would fill many entire issues of this journal. After-the-fact loss of initial control for any of the preceding reasons is common, not rare, in community program outcome research. That is one reason why not all program evaluation studies *reported*—and even fewer of those *done*—ever "mess with" controls. Of those that do, a substantial number, present company included, have been victimized either by incomplete or far less than ideal initial control, or by unavoidable breakdown of control during the study.

Faced with such natural disasters, investigators who care about control have several compensatory options to pursue. One is to bail out, as best one can, statistically. Analysis of covariance is a generic procedure designed to minimize initial mismatches and to bring comparison groups back to approximately the same starting point. Another way to deal with initial mismatches is to hack and chisel at the subject groups in hopes of paring them down to approximately comparable samples. Among the dangers of this procedure is that the adjustments may have to be asymmetrical (either because the original disproportionality comes more from one group than from the other or because the N is more robust in one group than the other). If the reduction in N comes primarily from one group, it may distort the group's defining characteristics. Moreover, such a procedure often highlights an inherent conflict between the ideal of a tight match (which, necessarily entails loss of N) versus the robustness and representativeness of the samples retained. Since this conflict is real, it is sometimes resolved by tolerating noise in the match to preserve sufficiently large and representative N s for the major substantive program evaluation analyses.

An interesting but more subtle question that the researcher sometimes faces is, When is a control not a control? Thus, groups can be well matched statistically but not "psychologically." Take as an example a program in which crisis interventionists are trained to

use special abreactive techniques in which they have interest but no special investment. The experimental question is whether the use of such techniques improves intervention outcomes with people experiencing current life crises. The groups seen by the specially trained and "regular" workers are well matched demographically and in terms of the nature and seriousness of the crises that they have experienced. The study's criteria include the clinician's preratings and postratings of a series of relevant patient behaviors. In such a situation, experimental and control interventionists may have different cognitive sets about the study, with experimentals thinking that "they are evaluating the effectiveness of this new *program*" and controls believing that "they are evaluating *my* effectiveness as a clinician." If such differential program views exist, experimentals' postprogram ratings are more likely either to be objective or to reflect personal (pro or con) views of the *program*, whereas controls, who see *themselves* as the focus of the study, are more likely to provide positive change estimates for clients. Should that happen, genuine program effects are obscured or lost. A similar example can be cited in evaluating the effectiveness of school-based intervention programs. Teachers of experimental children in such programs are more likely than those of controls to see their behavioral evaluations of children as program-related. Hence, their judgments may be influenced by their views of, and attitudes toward, the program. By contrast, teachers of control children, lacking a program metric, are more likely to see the rating task in the context of how good a job *I* have personally done with Johnny or Mary this year. If so, there is a pull for them to give more positive end-of-year ratings (Cowen et al., 1974).

Standard statistical controls may also not suffice in situations in which a program's main content and activities happen, incidentally, to involve a major structural change in the everyday lives of the target subjects. Assume, for example, that college student volunteers are trained to work with chronic hospitalized patients using rational-emotive techniques. Is this sufficient in evaluating the effectiveness of the intervention to have a matched experimental and control group? Probably not! Such a de-

might confound the program's ostensibly active ingredient (i.e., the therapeutic approach) with the fact that the procedure, intentionally, involved establishing a meaningful interpersonal relationship with people whose lives ordinarily lacked such relationships. An ideal third group that would strengthen the study's interpretive base, is one that could control for the personal relationship (e.g., with games and recreational activities) in the absence of an intentional, therapy-system thrust. Similar examples can be identified for selective therapeutic programs with adolescents in correctional settings, children in institutions for the retarded, or geriatric patients.

The unusual complexity of certain community settings (e.g., hospitals and schools) plus the fact that many are veritable laboratories for exploring many, ever-changing, program variations underscores the fact that mere demographic-statistical comparability of experiments and controls does not automatically solve the control problem. From an experimenter's standpoint, evaluation of a specific program would be clearest if there were *no* other special programs in either the experimental or control settings. Indeed, experimenters' special blinders may impel them to see the world that way, even though that view does not correspond, ecologically, to reality. More often than not, settings such as schools and hospitals house a variety of programs—formal and informal and short- and long-lived. Some of these programs may address the same objectives and behaviors as the experimental program in question (Freeman & Sherwood, 1969). A behavior modification program for hospital patients may take place alongside of drug-therapy and patient ward-governance programs. A school mental health intervention may co-occur with Glasser circle and Distar programs (Cowen et al., 1974). The intermixing of such programs not only makes it difficult to evaluate their separate contributions but also often means that an ostensibly pure experimental program is in fact that program plus several overlapping services or programs in one setting, compared to another (so-called control) setting, which happens not to have that particular program but does have three or four other programs addressed to similar

behaviors in comparable target subjects. Sometimes, in fact, an administrative decision is made to assign the special program to one setting *because* (compared to other similar ones) it is deficient in the type of services that the program provides. Conversely, control settings may be assigned other similar programs as part of an (understandable) administrative philosophy of sharing the wealth. Practical problems of control, in such situations, are magnified by the facts that some of the overlapping programs, either in experimental or control settings or both, are likely to be short-lived or to change in the process, and new programs may be introduced while the experiment is in progress. Although each of the foregoing possibilities is regrettable experimentally, they are part of the community's reality.

Problems of proper control, in community research, are diabolically complex and create serious persistent stumbling blocks to sound program evaluation research.

Overview and Summary

Communities are many things. One thing they are not is an ideal laboratory for anti-septic psychological studies. Their extraordinary complexity, omnipresent flux, action-service orientation, and susceptibility to day-to-day pressures present real and formidable barriers to "Mr. Clean" program evaluation studies. These factors place major constraints on the design of studies, the types of criteria that can be used, and the rigor of sophistication of the control that can be exercised. Although some of those problems can be reduced through judicious planning, others, quite beyond the experimenter's control, cannot. This is one reason why theory, logic, and the actual development and implementation of new community programs have outpaced the field's supporting research base.

The tugs and pulls of this situation are clear. On the one side is the obvious need to pose important, socially significant questions and to understand the impact and value of innovative practices designed to overcome long-standing, refractory problems in mental health. On the other are our training and bloodlines as experimenters and our un-

derstandings of past accepted canons for accreting new knowledge. These opposing tensions are as apparent in community research as in any subdomain of psychology today.

The intent of this article is not to discourage trying harder. Such effort is sorely needed; it can have great payoff value. Much can be done to strengthen community program evaluation technology and to design studies that reduce sources of confound or error. Weaknesses in specific measures or in classes of criteria typically used in community program outcome research dictate that greater emphasis be placed on converging sources of evidence. But we must still expect that community realities will remain to militate against ideal research studies. The vulnerability of findings from any single community evaluation study points to the importance both of replication and of tolerance for a slow accretive process, in which small pieces in a puzzle gradually cumulate toward weight-of-evidence conclusions about major new programming approaches. Although such a process is not intrinsically inimical to the way of science, it may be more caricatured in community research than in other fields.

The compelling logic of the community approach, the significance of the problems it addresses, and the excitement and clinical promise of some of its early innovative programmatic efforts have been sufficient to carry the field's infancy and early childhood. The future, however, will stand or fall on the solidity of its empirical footing. Social significance cannot, in that process, be sacrificed at the altar of laboratory precision. Hence, we must expect that successive approximations—the gradual putting together of sometimes chipped or scarred building blocks—will be the way of community program evaluation research in the coming decades.

For the reader who seeks wisdom and sophistication beyond the frailties of the present account, the following additional sources are suggested: Schulberg, Sheldon, and Baker, 1969; Bloom, 1972; Roen, 1971; Glass, 1976; Hammer, Landsberg, and Neigher, 1976; Fairweather and Tornatzky, 1977; Neigher, Hammer, and Landsberg, 1977; and Guttentag and Saar, 1978.

References

- Bloom, B. L. Mental health program evaluation. In S. E. Golann & C. Eisdorfer (Eds.), *Handbook of community mental health*. New York: Appleton-Century-Crofts, 1972.
- Bloom, B. L. Evaluating achievable objectives for primary prevention. In D. C. Klein & S. E. Goldston (Eds.), *Primary prevention: An idea whose time has come* (Department of Health, Education, and Welfare Publication No. (ADM) 77-447). Washington, D.C.: U.S. Government Printing Office, 1977.
- Brunswick, E. *Systematic and representative design of psychological experiments*. Berkeley, Calif.: University of California Press, 1947.
- Campbell, D. T. Factors relevant to the validity of experiments in social settings. In H. C. Schulberg, A. Sheldon, & F. Baker (Eds.), *Program evaluations in the mental health fields*. New York: Behavioral Publications, 1969.
- Cowen, E. L. Social and community interventions. *Annual Review of Psychology*, 1973, 24, 423-473.
- Cowen, E. L. Baby-steps toward primary prevention. *American Journal of Community Psychology*, 1971, 5, 1-22.
- Cowen, E. L., Lorton, R. P., & Dorr, D. Research in the community cauldron: A case report. *Canadian Psychologist*, 1974, 15, 313-325.
- Fairweather, G. W., & Tornatzky, L. G. *Experimental methods for social policy research*. New York: Pergamon Press, 1977.
- Freeman, H. E., & Sherwood, C. C. Research in large scale intervention programs. In H. C. Schulberg, A. Sheldon, & F. Baker (Eds.), *Program evaluations in the mental health fields*. New York: Behavioral Publications, 1969.
- Glass, G. V. (Ed.). *Evaluations studies review annual* (Vol. 1). Beverly Hills, Calif.: Sage Publications, 1976.
- Glidewell, J. C., Gildea, M. C.-L., & Kaufman, M. K. The preventive and therapeutic effects of two school mental health programs. *American Journal of Community Psychology*, 1973, 1, 295-329.
- Guttentag, N., & Saar, S. (Eds.). *Evaluations studies review annual* (Vol. 2). Beverly Hills, Calif.: Sage Publications, 1978.
- Hammer, R. J., Landsberg, G., & Neigher, W. (Eds.). *Program evaluation in community mental health centers*. New York: D and O Press, 1976.
- Hammond, K. R. Representative vs. systematic design in clinical psychology. *Psychological Bulletin*, 1954, 51, 150-159.
- Heller, K., & Monahan, J. *Psychology and community change*. Homewood, Ill.: Dorsey Press, 1977.
- Hereford, C. F. *Changing parental attitudes through group discussion*. Austin: University of Texas Press, 1963.
- Kelly, J. G. The quest for valid preventive interventions. In G. Rosenblum (Ed.), *Issues in community psychology and preventive mental health*. New York: Behavioral Publications, 1971.

- Neigher, W., Hammer, R. J., & Landsberg, G. (Eds.). *Emerging developments in mental health program evaluation*. New York: Argold Press, 1977.
- Roen, S. R. Evaluative research and community mental health. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavioral change: An empirical analysis*. New York: Wiley, 1971.
- Schulberg, H. C., Sheldon, A., & Baker, F. (Eds.).

Program evaluation in the mental health fields. New York: Behavioral Publications, 1969.

- Zax, M., & Klein, A. Measurement of personality and behavior changes following psychotherapy. *Psychological Bulletin*, 1960, 57, 435-448.
- Zdep, S. M., & Irvine, S. H. Reverse Hawthorne effect in educational evaluation. *Journal of School Psychology*, 1970, 8, 85-95.

Received December 31, 1977 ■

Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology

Paul E. Meehl
University of Minnesota

Theories in "soft" areas of psychology lack the cumulative character of scientific knowledge. They tend neither to be refuted nor corroborated, but instead merely fade away as people lose interest. Even though intrinsic subject matter difficulties (20 listed) contribute to this, the excessive reliance on significance testing is partly responsible, being a poor way of doing science. Karl Popper's approach, with modifications, would be prophylactic. Since the null hypothesis is quasi-always false, tables summarizing research in terms of patterns of "significant differences" are little more than complex, causally uninterpretable outcomes of statistical power functions. Multiple paths to estimating numerical point values ("consistency tests") are better, even if approximate with rough tolerances; and lacking this, ranges, orderings, second-order differences, curve peaks and valleys, and function forms should be used. Such methods are usual in developed sciences that seldom report statistical significance. Consistency tests of a conjectural taxometric model yielded 94% success with zero false negatives.

I had supposed that the title gave an easy tipoff to my topic, but some puzzled reactions by my Minnesota colleagues show otherwise, which heartens me because it suggests that what I am about to say is not trivial and universally known. The two knights are Sir Karl Raimund Popper (1959, 1962, 1972; Schilpp, 1974) and Sir Ronald Aylmer Fisher (1956, 1966, 1967), whose respective emphases on subjecting scientific theories to grave danger of refutation (that's Sir Karl) and major reliance on tests of statistical significance (that's Sir Ronald) are, at least in current practice, not well integrated—perhaps even incompatible. If you have not been accus-

tomed to thinking about this incoherency, and my remarks lead you to do so (whether or not you end up agreeing with me), this article will have served its scholarly function.

I consider it unnecessary to persuade you that most so-called "theories" in the soft areas of psychology (clinical, counseling, social, personality, community, and school psychology) are scientifically unimpressive and technologically worthless. Documenting that statement would of course require a considerable amount of time, but you can quickly get the flavor by having a look at Braun (1966); Fiske (1974); Gergen (1973); Hogan, DeSoto, and Solano (1977); McGuire (1973); Meehl (1960/1973a, 1959/1973f); Mischel (1977); Schlenker (1974); Smith (1973); and Wiggins (1973). These are merely some high visible and forceful samples; I make no claim to bibliographic completeness on the large theme of "What's wrong with 'soft' psychology." A beautiful hatchet job, which in my opinion should be required reading for all PhD candidates, is by the sociologist Andreski (1972). Perhaps the easiest way to convince yourself is by scanning the literature of soft psychology over the last 30 years and noticing what

This article is based on a lecture delivered at the meeting of the American Psychological Association, Washington, D.C., September 1976, on the occasion of the author's receiving Division 12, Section 3's Distinguished Scientist Award. The research reported here was assisted by Grant MH 24224 from the National Institute of Mental Health and the University of Minnesota Psychiatry Research Fund. Completion of the article was aided by a James McKeen Cattell Fund sabbatical award.

Requests for reprints should be sent to Paul E. Meehl, Mayo Box 392, University of Minnesota, Minneapolis, Minnesota 55455.

happens to theories. Most of them suffer the fate that General MacArthur ascribed to old generals—They never die, they just slowly fade away. In the developed sciences, theories tend either to become widely accepted and built into the larger edifice of well-tested human knowledge or else they suffer destruction in the face of recalcitrant facts and are abandoned, perhaps regretfully as a “nice try.” But in fields like personology and social psychology, this seems not to happen. There is a period of enthusiasm about a new theory, a period of attempted application to several fact domains, a period of disillusionment as the negative data come in, a growing bafflement about inconsistent and unreplicable empirical results, multiple resort to ad hoc excuses, and then finally people just sort of lose interest in the thing and pursue other endeavors.

Since I do not want to step on toes lest my propaganda falls on deaf ears, I dare not mention what strike me as the most egregious contemporary examples, so let us go back to the late 1930s and early 1940s when I was a student. In those days we were talking about level of aspiration. You could not pick up a psychological journal—even the *Journal of Experimental Psychology*—without finding at least one and sometimes several articles on level of aspiration in schizophrenics, or in juvenile delinquents, or in Phi Beta Kappas, or whatever. It was supposed to be a great powerful theoretical construct that would explain all kinds of things about the human mind from psychopathology to politics. What happened to it? Well, I have looked into some of the recent textbooks of general psychology and have found that either they do not mention it at all—the very phrase is missing from the index—or if they do, it gets cursory treatment in a couple of sentences. There is no doubt something to the notion. We all agree (from common sense) that people differ in what they demand or expect of themselves, and that this probably has something to do, sometimes, with their performance. But it did not get integrated into the total nomological network, nor did it get clearly liquidated as a nothing concept. It did not get killed or resurrected or transformed or solidified; it just kind of dried up and blew away, and we no longer wanted to talk about it or do experi-

mental research on it. A more recent example is the theory of “risky shift,” about which Cartwright (1973) wrote, after reviewing 196 papers that appeared in the 1960s:

As time went by . . . it gradually became clear that the cumulative impact of these findings was quite different from what had been expected by those who produced them. Instead of providing an explanation of why “groups are riskier than individuals,” they in fact cast serious doubt on the validity of the proposition itself (p. 225).

It is now evident that the persistent search for an explanation of “the risky shift” was misdirected and that any adequate theory will have to account for a much more complicated set of data than originally anticipated. But it is not clear how theorizing should proceed, since serious questions have been raised as to whether, or in what way, “risk” is involved in the effects to be explained (p. 226).

After 10 years of research, [the] original problem remains unsolved. We still do not know how the risk-taking behavior of “real-life” groups compares with that of individuals (p. 231).

I do not think that there is any dispute about this matter among psychologists familiar with the history of the other sciences. It is simply a sad fact that in soft psychology theories rise and decline, come and go, more as a function of baffled boredom than anything else; and the enterprise shows a disturbing absence of that *cumulative* character that is so impressive in disciplines like astronomy, molecular biology, and genetics.

There are some solid substantive reasons for this that I will list here, lest you think that I am beating up on the profession, unaware of the terrible intrinsic difficulty of our subject matter. Since (in 10 minutes of superficial thought) I easily came up with 20 features that make human psychology hard to scientize, I invite you to pick your own favorites. Differences as to which difficulties are emphasized will not, I am sure, cause any disagreement about the general fact. This is not the place to develop in detail the thesis that the human mind is hard to scientize, let alone to prove it. Each of the 20 difficulties is, I am aware, debatable; and one could find competent psychologists who would either deny a difficulty's reality—at least in the form I state it—or who, although admitting it exists, would maintain that we have, or will be able

to develop shortly, methods adequate to overcome or circumvent it. Each of these alleged difficulties in scientizing the human mind is sufficiently controversial to deserve a methodological article by itself. This being so, to substitute a once-over lightly (and hence inevitably dogmatic) defense of each as a real difficulty is, for those who accept it, a work of supererogation, and for the others, it is doomed to failure. I therefore confine myself to listing and explaining the problems, repeating that my purpose in so doing is to prevent the rest of my article from being taken as a kind of malicious and unsympathetic attack on psychologists (of which, after all, I *am* one!) based on an inadequate appreciation of the terrible difficulties under which we work. In a few cases I have explained at some length and replied to objections, these being cases in which a difficulty is not widely recognized in our profession or in which it is generally held to have been disposed of by a familiar (but erroneous) refutation or solution. Regrettably, some psychologists use "philosophical" arguments that are a generation or more out of date.

Since I am listing and summarizing rather than developing or proving, it seems appropriate to present the set of difficulties as follows:

1. *Response-Class Problem*

This involves the well-known difficulties of slicing up the raw behavioral flux into meaningful intervals identified by causally relevant attributes on the response side, a problem that exists already in the Skinner box (Skinner, 1938, p. 70), worsens in field study by an ethologist, and reaches almost unmanageable proportions in studying human social behavior of the kind to which clinical, social, and personality psychologists must address themselves (see, e.g., MacCorquodale & Meehl, 1954, pp. 218-231, after a quarter century still considered by some as best statement of the problem; Hinde, 1970, pp. 10-13; Meehl, 1954, pp. 40-44 and chap. 6 *passim*; Skinner, 1938, pp. 33-43).

2. *Situation-Taxonomy Problem*

As is well-known, the importance of an adequate classification and sampling of environ-

ments and situations has received less attention than Problem 1, above, despite emphasis by several major contributors such as Roger Barker (1968), Egon Brunswik (1955), and Saul B. Sells (1963). It seems likely that the problems of characterizing the stimulus side, even though often neglected by the profession or dealt with superficially, are about as intractable as the characterization of the response class. It is not even clear whether identification and measurement of the relevant stimulus dimensions (e.g., size) is the same task as concocting a taxonomy of "situations" and "environments," nor whether the answer to this question would quickly generate rules for an adequate statistical ecology applicable to research design. So I am perhaps lumping under this "situation-taxonomy" rubric three distinguishable but related problems. I am inclined to think that most (not all) of the current methodological controversy concerning traits versus situations is logically and mathematically reducible to this and the preceding category, since I think that traits are disposition clusters, and dispositions always involve at least implicit reference to the stimulus side; but this is not the place to push that view.

3. *Unit of Measurement*

One sometimes hears this conflated with one or both of the preceding, but, of course, it is not the same. There are questions in rating scales and in psychometrics (as well as in certain branches of nondifferential psychology) in which disagreements persist about such fundamental matters as the necessity of a genuine interval or ratio scale for the use of certain kinds of sampling statistical inference.

4. *Individual Differences*

Perhaps the shortest way to discuss this one is to point out the oddity that what is one psychologist's subject matter is another psychologist's error term (Cronbach, 1957)! More generally, the fact is that organisms differ not only with respect to the strengths of various dispositions, but, more common and more distressing for the researcher, they differ as to *how* their dispositions are shaped and

organized. As a result, the individual differences involved in "mental chemistry" are tougher to deal with than, say, the fact that different elements have different atomic numbers or that elements with the same atomic number vary in atomic weights (isotopes).

5. Polygenic Heredity

It is generally conceded that the measurement and causal inference problems that arise in biometrical genetics are, with some exceptions, more difficult than those found in the kind of single factor dominant or recessive gene situation on which the science of genetics was originally founded. Except for Mendelizing mental deficiencies and perhaps some psychiatric disorders that are transmitted in a Mendelizing fashion, most of the attributes studied by soft-field psychologists are influenced by polygenic systems. Usually we must assume that several totally different and unrelated polygenic systems influence a manifest trait like social introversion. Introversion may be based in part on a unitary (although polygenic) variable, as shown by Gottesman (1963) and others. However, as an acquired disposition of the adult-acclimated individual, it presumably results from a confluence of different polygenic contributors such as basic anxiety readiness, mesomorphic toughness, garden-variety social introversion, dominance, need for affiliation, and the like.

6. Divergent Causality

As pointed out 35 years ago by the physical chemist Irving Langmuir (1943; London, 1946; Meehl, 1954, pp. 60-61; Meehl, 1967/1970b, especially Footnotes 1-8 on pp. 395-396), there are complex systems whose causal structure and boundary conditions are such that slight differences—including those that are, for practical predictive and explanatory purposes, effectively "random" (whatever their inner deterministic nature may be)—tend to "wash out," "cancel each other," or "balance" over the long run. On the other hand, there are other systems in which such slight perturbations or differences in the exact character of the initial conditions are, so to speak, amplified over the long run. Langmuir christened

the former kind of causality as "convergent," as when we say that the average errors in making repeated measurements of a table length tend to cancel out and leave us with a stable and highly trustworthy mean value of the result. On the other hand, an object in unstable equilibrium can lean slightly toward the right instead of the left, as a result of which a deadly avalanche occurs burying a whole village. Although both sorts of systems are found at all levels of Comte's Pyramid of the Sciences, it seems regrettably true that the incidence of important and pervasive types of divergent causality is greater in the sciences of behavior.

7. Idiographic Problem

It is not necessary to "settle" the long-continued methodological controversies regarding idiographic versus nomothetic methods in psychology and history (e.g., whether they are philosophically, metaphysically fundamentally different) to agree with strong proponents of the idiographic method, such as Gordon Allport (Allport, 1937) or my long-time friendly adversary on the prediction issue, Robert R. Holt (1958), that the human personality—unless one approaches it with the postulate of impoverished reality—has in its content, structure, and, conceivably, even in individual differences as to some of its "laws," and very much in its origins, properties and relations that make the study of personality rather more similar to such disciplines as history, archeology (historical), geology, or the reconstruction of a criminal case from police evidence than the derivation of the molar gas laws from the kinetic theory of heat or the mechanisms of heredity from molecular biology. Some would argue that such explanatory derivations aside, even the mere inductive subsumption of particulars (episodes, molar traits, persons) under descriptive generalizations is a more difficult and problematic affair in these disciplines than in most branches of physical and biological science.

8. Unknown Critical Events

Related to divergent causality and idiographic understanding but distinguishable

from them is the fact that critical events in the history of personality development are frequently hard to ascertain. There is reason to believe that in some instances they are literally never ascertained by us or known to the individual under study, even somebody who has spent 500 hours on the analytic couch. They are sometimes observable events that, however, were not in fact observed and recorded, such as the precise tone of voice and facial expression that a patient's father had when he was reacting to an off-color joke that the patient innocently told at the dinner table at age 7. Every thoughtful clinician realizes that the standard life history that one finds in a medical chart is, from the standpoint of thorough causal comprehension, so thin and spotty and selective as to border on the ludicrous. But there is also what I would view as an important causal source of movement in one rather than another direction of divergent causality, namely, inner events, such as fantasies, resolutions, shifts in cognitive structure, that the patient may or may not report and that he or she may later be unable to recall.

9. Nuisance Variables

Other things equal, it is handy for research and theorizing if we can sort out the variables into three classes, namely, (a) variables that we manipulate (in the narrow sense of the word *experimental*), (b) variables that we do not manipulate but can hold constant or effectively exclude from influence by one or another means isolating the system under study, and (c) variables that are quasirandom with respect to the phenomena under study, so that they only contribute to measurement error or the standard deviation of a statistic. Unfortunately, there are systems, especially social and biological systems of the kind that clinical psychologists and personologists study, in which there is operative a nonnegligible class of variables that are not random but systematic, that exert a sizable influence, and are themselves also sizably influenced by other variables, either exogenous to the system (F. M. Fisher, 1966) or contained in it, such that we have to worry about the influence of these variables, but we cannot always ascertain the

direction of the causal arrow. Sometimes we cannot even get sufficiently trustworthy measurements of these variables so as to "partial out" or "correct" their influence even if we are willing to make conjectures about the direction of causality. There are some circumstances in which we can extrapolate from experimental studies or from well-corroborated theory to make a high-confidence decision about the direction of causal influence, but there are many other circumstances—in soft psychology, the preponderating ones—in which this is not possible. Further, lacking special configurations such as highly atypical cells in a multivariate space or correlation coefficients that impose strong constraints on a causal interpretation, or provisional assumptions as relied on in path analysis (Li, 1975), the system is statistically and causally indeterminate. (Why these constraints are regularly treated as "assumptions" instead of refutable conjectures is itself a deep and fascinating question that I plan to examine some other time.) The well-known difficulties in assessing the influence of socioeconomic status (SES) on children's IQ when unscrambling the hereditary and environmental contributors to intelligence is perhaps the most dramatic one, but other less emotion-laden examples can be found on all sides in the behavioral sciences. (See Meehl, 1970a, 1971/1973b).

10. Feedback Loops

A special case in engineering is the usual in psychology, that a person's behavior affects the behavior of other persons and hence alters the schedule imposed by the "social Skinner box." The complexities here are so refractory to quantitative decomposition that yoked box setups came to be used even for the (relatively simple) animal case as a factual substitute for piecewise causal-dispositional analysis. In the human social case, they may be devastating.

11. Autocatalytic Processes

The chemist is familiar under the label *autocatalysis* with a rare but important kind of preparation in which one of the end products of the chemical processes is itself capable of catalyzing the process. Numerous common ex-

examples spring to mind in psychology, such as anxiety and depression as affects or economic failure as a social impact. Much of neurosis is autocatalytic in the cognitive-affective-volitional system, as are counterneurotic healing processes. When this kind of complicated setup is conjoined with the critical event, idiographic, and divergent causality factors, and also with the individual differences factor (that parameters relating the growth of one state of schedule to a dependent variable, which itself in turn acts autocatalytically, show individual differences), the task of unscrambling such a situation becomes terribly difficult.

12. *Random Walk*

There is a widespread and understandable tendency to assume that the class of less-probable outcomes, given constancy of other classes of causally efficacious variables, should in principle be explicable by detecting a class of systematic input differences. Thus, for instance, we try to understand the genetic/environmental contributions to schizophrenia by studying discordant monozygotic twins. If I develop a florid clinical schizophrenia and my monozygotic twin remains sane and wins the Pulitzer Prize for poetry, it is a sensible strategy for the psychologist to consider my case *and similar cases* with an eye for "systematic differences" (such as who was born first, who was in what position in the uterus, or who had a severe case of scarlet fever with delirium) as responsible for dramatic difference in final outcomes. When one reflects on the rather meager yield of such assiduous ferretting out of systematic differences by, say Gottesman and Shields (1972) in their excellent book, one experiences bafflement. On the one hand, the concordance rate for monozygotic twins is only a little over 50%, indicating a very large nongenetic component in causality. Yet, on the other hand, we find feeble or null differences when we look at the list of "obvious, plausible" differentiators between the twins who fall ill and the twins who remain well. Of course, one can always say—and would no doubt be partly right in this—that we just have not been clever enough to hit on the right ones; or even if, qualitatively, they are the right ones, we do not have sufficiently

construct-valid measures of them to show up in the statistics.

There is, however, an alternative explanation that when one reflects on it, is plausible (at least to a clinical practitioner like myself) and that has analogues in organic medicine and in other historical sciences like geology or the theory of evolution, to wit, that we are mistaken to look for a "big systematic variable" of the kind that is already in our standard list of influences, such as organic disease, parental preference, or SES of an adoptive home. Rather, we might emphasize that a human being's life history involves as one form of divergent causality, something akin to the stochastic process known as a "random walk" (Bartlett, 1955, pp. 15–20, 47–50, 89–96; Feller, 1957, pp. 73, 311; Kemeny, Snell, & Thompson, 1957, pp. 171–177; Read, 1972, pp. 779–782). At several points that are individually minor but collectively critical determinative, it is an almost "chance" affair whether the patient does A or not A, whether his girl friend says she will or will not go out with him on a certain evening, or whether he happens to hit it off with the ophthalmologist that he consults about some peculiar vision disturbances that are making him anxious about becoming blind, and the like. If one twin becomes psychotic at the end of such a random walk, it is possible that he was suffering from what was only, so to speak, "bad luck"—not a concept that appears in any standard list of biological and social nuisance variables!

Luck is one of the most important contributors to individual differences in human suffering, satisfaction, illness, achievement, and so forth, an embarrassingly "obvious" point that social scientists readily forget (Gunther, 1977; Jencks, 1972, pp. 8–9, 227–228; Popper, 1974, pp. 36–37; Stoddard, 1929; for further discussion of this see Meehl, 1972/1973g, pp. 402–407, Meehl, 1973d, pp. 220–221). Of course, the fact that a process resembles a random walk does not mean that it is not susceptible to quantitative treatment. Witness the extensive formal development of this sort of process in the field of finite mathematics by engineers and others. The point is that its analytical treatment will not look like the familiar kind of search for a systematic class

of differentiating variables like SES as a nuisance variable in relationship to educational outcome and intelligence.

13. *Sheer Number of Variables*

I suppose that this is the most commonly mentioned of the difficulties of social science, and I assume that my readers would accept it without further elaboration. But it is worth mention that the number of variables is large from several different viewpoints. Thus we deal on one side with a *large number of phenotypic traits*, conceiving a phenotypic trait as a related family of response dispositions that (a) are correlated to some stipulated degree pairwise and that (b) have some kind of logical, semantic, social, or other "meaning" overlap or resemblance that entitles us to class them together. Or, again, we consider a large number of dimensions on the stimulus side and on the response side that are relevant in formulating a law of behavior acquisition, as well as in the subsequent control and activation dispositions thus acquired. From still another viewpoint, the list of historical causal influences is long and heterogeneous, ranging from such diverse factors as a mutated gene or a never-diagnosed subclinical tuberculosis to a mother who mysteriously permitted herself the fantasy that a brutal father would go away, and the like. It should be noted that this matter of sheer number of variables would not be so important (except as a contributor to residual "random variation" in various kinds of outcomes) if they were each small contributors and independent, like the sources of error in the scattering of shots at a target in classical theory of errors. But in psychology this is not typically the situation. Rather, the variables, although large in number, are each nuisance variables that carry a significant amount of weight, interact with each other, and contribute to idiographic development via the divergent causal-ity mode.

14. *Importance of Cultural Factors*

This source of individual differences, both in acquired response clusters (traits) and in

the parameters of acquisition and activation functions, especially when taken together with the genetic factors contributing, for instance, to social competence, mental health, intellect, and so on, makes for unusual complications in understanding how somebody got to be the way he is. We are, for instance, so accustomed to referring to nuisance variables like SES in considering the design of experiments that involve SES-related individual differences that we readily forget something every reflective person knows—that the measures of things like SES are general and not tailor-made for what is idiographically more significant in the development of a particular person. So when we speak of "controlling for SES," that is a loose use of language in comparison with "controlling the temperature" in a Skinner box or controlling the efflux of calories in a physics lab by use of a bomb calorimeter. A treatise on the principles of internal medicine (such as Harrison et al., 1966) sometimes refers to cultural factors, including those that are not at all understood—in the way that, say, dietary deficiency might be mediated by extreme poverty in a backward country—and simply says that for some reason this disease is found more frequently among the rich than among the poor. But the *important* causal chains of prime interest to the physician, even in his role as an advisor of preventive medicine, do not typically involve worry about whether somebody is fifth-generation upper class or the third child of parents who became anxious after the birth of the second oldest sibling. However, this kind of consideration might be crucial in reconstructing the life history of such a person.

15. *Context-Dependent Stochastologicals*

Cronbach and Meehl (1955/1973) and subsequent writers adopted (from the neopositivist philosophers of science) the phrase *nomological network* to designate the system of lawlike relationships conjectured to hold between theoretical entities (states, structures, events, dispositions) and between theoretical entities and their observable indicators. The "network" metaphor is chosen to emphasize the structure of such systems, in which the *nodes* of the network, representing the postu-

lated theoretical entities, are connected by the *strands* of the network, representing the lawful relationships hypothesized to hold between the entities. What makes such a set of theoretical statements a system (rather than a mere conjunction of unrelated assertions, a "heap of hypotheses") is the semantic fact of their shared terms, an overlap in the propositions' inner components, without which, of course, no deductive fertility and no derivation chains to observational statements would be formally possible. The network is empirical (and "scientifically respectable"), because a proper subset of the theoretical terms is co-ordinated in fairly direct ways ("operationally") with terms designating perceptual or instrument-reading predicates. These latter predicates normally possess the admirable properties of *quick decision*, *minimal theory dependence*, and *high interpersonal consensus*.

Despite the current distaste for these "objectivist" conceptions, I remain an old-fashioned unreconstructed positivist to the limited extent that I think science—both "normal science" and "revolutionary, paradigm-replacing science"—differs from less promising, non-cumulative, and personalistic enterprises like politics, psychotherapy, folklore, ethics, metaphysics, aesthetics, and theology *in part* because of its skeptical insistence on reliable (intersubjective, replicable) protocols that describe observations. Skinner is in better shape than Freud partly because Norman Campbell (1920/1957, p. 29) was right in saying that the kinds of judgments for which universal assent can be obtained are (a) judgments of temporal simultaneity, consecutiveness, and "betweenness"; (b) judgments of coincidence and "betweenness" in space; and (c) judgments of number. I cannot view the increasingly fashionable dismissal of these objectivity-oriented views as other than obscurantist in tendency. (See Kordig, 1971, 1973.)

However, the nomological network, even though correlated directly, here and there, with observational data, is not "operational" throughout, since some of the nodes and strands are connected with the observational data base only via other subregions of the network. As Hempel said (1952):

A scientific theory might therefore be likened to a complex spatial network: Its terms are represented

by the knots, while the threads connecting the latter correspond, in part, to the definitions and, in part to the fundamental and derivative hypotheses included in the theory. The whole system floats, as it were, above the plane of observation and is anchored to it by rules of interpretation. These might be viewed as strings which are not part of the network but link certain points of the latter with specific places in the plane of observation. By virtue of those interpretive connections, the network can function as a scientific theory: From certain observational data, we may ascend, via an interpretive string, to some point in the theoretical network, thence proceed, via definitions and hypotheses, to other points, from which another interpretive string permits a descent to the plane of observation. (p. 36)

Even though the core of these ideas is sound and important, the word *nomological* is in soft psychology at best an extension of meaning and at worst a misleading corruption of the logician's terminology. Originally it designated strict laws as in W. E. Johnson's (1921/1964) earlier use of "nomic necessity" (p. 61). The lawlike relationships we have to work with in soft psychology are rarely (never?) of this strict kind, errors of measurement aside. Instead, they are correlations, tendencies, statistical clusterings, increments of probabilities, and altered stochastic dispositions. The ugly neologism *stochastological* (as analogue to *nomological*) is at least shorter than the usual "probabilistic relation" or "statistical dependence," so I shall adopt it. We are so accustomed to our immersion in a sea of stochastologicals that we may fail to notice what a terrible disadvantage this sort of probabilistic law network puts us under, both as to the clarity of our concepts and, more importantly, the testability of our theories. (One still hears the tiresome complaint that a theoretical system cannot be simultaneously concept definatory and factually assertive, despite repeated explanations of how this works. See, e.g., Braithwaite, 1960, pp. 76-87; Campbell, 1920/1957, pp. 119-158; Carnap, 1936-1937/1950, 1952/1956, 1966, pp. 225-226, 265-274; Feigl, 1956, pp. 17-19; Hempel, 1952, 1958, pp. 81-87; Lewis, 1970; Maxwell, 1961, 1962; Meehl, 1977, pp. 35-37; Nagel, 1961, pp. 87, 91-93; Pap, 1958, pp. 318-321, 1962, pp. 46-52; Popper, 1974, pp. 14-73; Ramsey, 1931/1960; Sellars, 1948.)

When the observational corroborators of the theory consist wholly of percentages, crude

curve fits, correlations, significance tests, and distribution overlaps, it is difficult or impossible to see clearly when a given batch of empirical data refutes a theory or even when two batches of data are (in any interesting sense) "inconsistent." All we can usually say with quasi-certainty is that context-dependent statistics should *not* be numerically identical in different studies of the same problem. (A dramatic recent example of this was the discovery that some of Sir Cyril Burt's correlation coefficients were *too consistent* to have been derived from the different tests and populations that he reported!)

In heading this section "Context-Dependent Stochastologicals," I mean to emphasize the aspect of this problem that seems to me most frustrating to our theoretical interests, namely, that the statistical dependencies we observe are always somewhat, and often strongly, dependent on the institution-cum-population setting in which the measurements were obtained. Lacking a "complete (causal) theory" of what influences what, *and how much*, we simply cannot compute expected numerical changes in stochastic dependencies when moving from one population or setting to another. Sometimes we cannot even rationally predict the direction of such changes. If the difference between two Pearson correlations were safely attributable to random sampling fluctuation alone, we could use the statistician's standard tools to decide whether Jones's study "fails to replicate" Smith's. But the usual situation is not one of simple cross-validation shrinkage (or "boostage")—rather, it involves the validity generalization problem. For this, there are no standard statistical procedures. We may be able, relying on strong theorems in general statistics plus a backlog of previous experience and a smattering of theory, to say some fairly safe things about restriction of range and the like. However, thoughtful theorists realize how little *quantitatively* we can say with sufficient confidence to warrant counting an unexpected shift in a stochastic quantity as a strong "discorroborator." This being so, we cannot fairly count an "in the ball park" predicted value as a strong corroborator. For example, Meehl's Mental Measure correlates .50 with SES in Duluth junior high school students, as predicted from Fis-

bee's theory of sociability. When Jones tries to replicate the finding on Chicano seniors in Tucson, he gets $r = .34$. Who can say anything theoretically cogent about this difference? Does any sane psychologist believe that one can do much more than shrug?

Although probability concepts (in the theory) and statistical distributions (in the data) sometimes appear in both classical and quantum physics, their usual rôle differs from that of context-dependent stochastologicals in social science. Without exceeding space limitations or my competence, let me briefly suggest some differences. When probabilities appear in physics and chemistry, they often drop out in the course of the derivation chain, yielding a quasi-nomological at its termination (e.g., derivation of gas laws or Graham's diffusion law from the kinetic theory of heat, in which the postulates are nomological, the "conditions" are probability distributions, and the resulting theorems are again nomological). Second, when the predicted observational result still contains statistical notions, their numerical values are either not context dependent or the context dependencies permit precise experimental manipulation. A statistical scatter function for photons or electrons can be finely tuned by altering a very limited number of experimental variables (e.g., wavelength, slit width, screen distance), and the law of large numbers assures that the expected "probabilistic" values of, say, photon incidence in a specified band will be indiscernibly different from the observed (finite but huge) numbers.

All this is very unlike the stochastologicals of soft psychology, in which strong context dependence prevails, but we do not know (a) the complete list of contextual influences, (b) the function form of context dependency for those influences that we can list, (c) the numerical values of parameters in those function forms that we know or guess, or (d) the values of the context variables if we are so fortunate as to get past Ignorances a-c. Finally, unlike physics, our sample sizes are usually such that the Bernoulli theorem does not guarantee a close fit between theoretical and observed frequencies—perhaps one of the few good uses for significance tests?

16. Open Concepts

As a consequence of the factors listed supra, especially those numbered 4, 7, 9, 15, it is usually not possible in the soft areas of social science to provide rigorous, explicit, or—the holy word when I was in graduate school—operational definitions for theoretical concepts. This difficulty occurs not because psychologists are intellectually lazy or sloppy, although most of us are at times (some routinely and on principle). Rather, it arises from the intrinsic nature of the subject matter, that is, from the organism's real compositional nature and structure and the causal texture of its environment. As has often been pointed out, one can concoct quick and easy "operational definitions" of psychological terms, but they will usually lack theoretical interest and, except for some important special cases (e.g., purely predictive task-tailored psychometrics and some kinds of operant behavior control), generalizable technological power (Lazarus, 1971; Loevinger, 1957). It is remarkable evidence of cultural lag in intellectual life that one can still find quite a few psychologists who are hooked on the dire necessity of strictly operational definitions, and who view open concepts as somehow methodologically sinful, although it is now a quarter of a century since the late Arthur Pap published his brilliant article on open concepts (Pap, 1953, see also chap. 11 of Pap, 1958). To do justice, and highlight the cultural lag, I should mention the related article of Waismann that antedated Pap's by 8 years (Waismann, 1945) and even Carnap's of 40 years ago (1936–1937/1950). I cannot name a single logician or a philosopher (or historian) of science who today defends strict operationism in the sense that some psychologists claim to believe in it. (They don't really—but you have to listen awhile to catch the deviations in *substance* when pseudooperationists are not discoursing dogmatically about *method*.)

The problem of open concepts and their relation to empirical falsifiability warrants a separate article, with which I am currently engaged, but suffice it to say here that the unavailability of open concepts in social and biological science tempts us to sidestep it by fake operationism on the one side (if we are

of the tough-minded, superscientific orientation) or to be contented with fuzzy verbalisms on the other side (if we are more artsy-craftsy or literary), thinking that it is the best we can get. The important point for methodology of psychology is that just as in statistics one can have a *reasonably precise theory of probable inference*, being "quasi-exact about the inherently inexact," so psychologists should learn to be sophisticated and rigorous in their metathinking about open concepts at the substantive level. I do not mean to suggest in saying this that the logicians' theory of open concepts is in a highly developed state, but it is far more developed than one would think from reading or listening to most psychologists.

I have elsewhere (Meehl, 1977) distinguished three kinds of openness that are involved in varying degrees in various psychological concepts and that may all be present in the same theoretical construct, namely, (a) openness arising from the indefinite extensibility of our provisional list of operational indicators of the construct; (b) openness associated with each indicator singly, because of the empirical fact that indicators are only probabilistically, rather than nomologically, linked to the inferred theoretical construct; and (c) openness due to the fact that most of our theoretical entities are introduced by an implicit or contextual definition, that is, by their role in the accepted nomological network, rather than by their inner nature. By their "inner nature" I mean nothing spooky or metaphysical but merely their ontological structure or composition as the latter will, with the progress of research, be formulatable in terms of the theoretical entities of more basic sciences in Comte's pyramid. In social and biological science, one should keep in mind that *explicit definition* of theoretical entities is seldom achieved in terms of the initial observational variables of those sciences, but it becomes possible instead by theoretical reduction or fusion. Explicit definition is achieved, if ever, in terms of some more basic underlying science (Meehl, 1977, see also Cronbach & Meehl (1955/1973); Meehl, 1959/1973f, 1973h, pp. 285–288).

A final remark, which also deserves fuller treatment in another place, is that when we

deal with open concepts, as in personality psychometrics of traits or taxa, the statistical phenomenon of *psychometric drift* as a result of bootstrap operations, refinement of measures, and theoretical reflection on the big matrix of convergent and discriminative validities (Campbell & Fiske, 1959) also generates, via our reliance on implicit or contextual definitions of theoretical entities, an associated *conceptual drift*, a meaning shift. When we reassign weight to fallible indicators of an entity to the extent that the very meaning of the term designating that entity is specified by its role in the network, such reassignment of weights—especially under drastic revisions of the system such as dropping a previously relied-upon indicator—constitutes a change in the theoretical concept. Difficult interpretative and research strategy problems arise here, because, on the one hand (especially in psychometrics) we encounter the danger that the resulting conceptual drift has pulled us away from what we started out to measure, but we also recognize that in psychology, as in the other sciences, part of the research aim is precisely that of bringing about revisions of concepts on the basis of revisions of the nomological network that implicitly defines them. We want, as Plato said, to carve nature at its joints; and the best test of this achievement is increased order in our material.

17. *Intentionality, Purpose, and Meaning*

We do not need to settle the philosopher's question of what is the essential condition for the existence of intentionality, nor buy Brentano's famous criterion that intentionality is the distinctive mark of the mental, to recognize that human beings think and plan and intend, that if rats do so they do it at a much lower level, that sunflowers probably do not, and that stones certainly do not. The formulation of powerful functional relationships for systems that do not possess the capacity to think, worry, regret, plan, and intend is obviously on the average an easier task. (But see Vico, 1744/1948, for a view so different that an American social scientist of our time can hardly grasp it.)

18. *Rule Governance*

Related to intentionality but sufficiently important to deserve a special listing is the fact that human behavior is rule governed. People do something not merely "in accordance with" a generalization but because they feel bound to obey the generalization stated in the form of a rule. Nobody has succeeded in coming up with a fully satisfactory definition of when a rule is a rule, but a sufficiently good approximation is to say that a rule differs from an empirical generalization in that a rule is not liquidated by being broken, whereas an empirical generalization is thereby liquidated (assuming that the conditions stated in its antecedent clause are granted, and the violation event is admitted into the corpus). Continued controversies in psycholinguistics reflect the importance of this kind of consideration in any discussion of human conduct.

19. *Uniquely Human Events and Powers*

In addition to being rule governed, there are several other human features that we do not share with chimpanzees, let alone sponges or boulders. I recall the late Richard M. Elliott saying that the main reason that psychology had done so poorly in its "theories" of humor is that man is the only animal that laughs. I think he had a good point here, since we have learned so much about aspects of human functioning, such as digestion and reproduction, by the experimental study of animals. There are a number of other things that human beings do that no infrahuman animal does, so far as we know. Only man speculates about nonpractical, theoretical matters; only man worships; only man systematically goes about seeking revenge, years later, for an injury done to him; only man carries on discussions about how to make decisions; and there are some features of cultural transmission that only man engages in, although the evidence now indicates that numerous other species transmit learned forms of behavior to subsequent generations.

20. *Ethical Constraints on Research*

This one is so obvious as to need no exposition. One can readily conceive quasi-definitive

experiments on the IQ-heredity controversy, whether there are family dynamics sufficient to make just anyone into a manic-depressive, that cannot be performed because to do so would be immoral.

Not to be overly pessimistic, let me mention (without proof) five noble traditions in clinical psychology that I believe have permanent merit and will still be with us 50 or 100 years from now, despite the usual changes. Some of these are currently unpopular among those addicted to one of the contemporary fly-by-night theories, but that does not bother me. These five noble traditions are (a) descriptive clinical psychiatry, (b) psychometric assessment, (c) behavior genetics, (d) behavior modification (I lump under this rubric positive contingency management, aversion therapy, and desensitization), and (e) psychodynamics. This list should convince you that I am not using methodological arguments to grind any substantive ax. I am probably one of the few psychologists alive today who would list all five of these as great, noble, and enduring intellectual traditions. I particularly emphasize the last, psychodynamics, since I am often perceived as a dust bowl empiricist who does not think that anything can be true or useful if it is not either based on laboratory experiments or statistical correlations. There is not a single experiment reported in my 23-volume set of the standard edition of Freud nor is there a *t* test. But I would take Freud's clinical observations over most people's *t* tests any time. I am confident that psychoanalytic concepts will be around after rubber band theory, transactional theory, attachment theory, labeling theory, dissonance theory, attribution theory, and so on, have subsided into a state of innocuous desuetude like the risky shift and level of aspiration. At the very least, psychoanalysis is an interesting theory, which is more than I can say about some of the "theories" that are currently fashionable.

These five noble traditions differ greatly in the methods they use and their central concepts, and I am hard put to say what is common among them. Some of them, such as behavior modification, are not conceptually exciting to those of us who are interested in ideas like Freud's, but they more than make up for that by their remarkable technological power.

I shall focus the remainder of my remarks on one feature that they have in common with the developed sciences (physical or biological); to wit, they were originally developed with negligible reliance on *statistical significance testing*. Even the psychometric assessment tradition in its early stages paid little attention to significance testing except (sometimes) for finding good items. Binet did not know anything about *t* tests, but he drew graphs of the developmental change of items. I suggest to you that Sir Ronald has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology.

It is easiest to see this from the methodological viewpoint of Sir Karl Popper, but fortunately we have here a rare instance in which Sir Karl's position yields the same result as the Bayesians', and both give the same result as "scientific common sense" practiced by those chemists and biologists who know nothing about philosophy of science or Bayesian statistics and could not care less about either. Briefly and simplistically, the position of Popper and the neo-Popperians is that we do not "induce" scientific theories by some kind of straightforward upward seepage from the clearly observed facts, nor do we "confirm" theories as the Vienna positivists supposed. All we can do is to subject theories—including the wildest and "unsupported" armchair conjectures (for a Popperian, completely kosher)—to grave danger of refutation, in accordance with the formally valid fourth figure of the implicative syllogism: $p \rightarrow q, \sim q, \therefore \sim p$, Popper's famous *modus tollens*.

A theory is corroborated to the extent that we have subjected it to such risky tests; the more dangerous tests it has survived, the better corroborated it is. If I tell you that Meehl's theory of climate predicts that it will rain sometime next April, and this turns out to be the case, you will not be much impressed with my "predictive success." Nor will you be impressed if I predict more rain in April

than in May, even showing three asterisks (for $p < .001$) in my t -test table! If I predict from my theory that it will rain on 7 of the 30 days of April, and it rains on exactly 7, you might perk up your ears a bit, but still you would be inclined to think of this as a "lucky coincidence." But suppose that I specify *which* 7 days in April it will rain and ring the bell; then you will start getting seriously interested in Meehl's meteorological conjectures. Finally, if I tell you that on April 4th it will rain 1.7 inches (.66 cm), and on April 9th, 2.3 inches (.90 cm) and so forth, and get seven of these correct within reasonable tolerance, you will begin to think that Meehl's theory must have a lot going for it. You may believe that Meehl's theory of the weather, like all theories, is, when taken literally, false, since probably all theories are false in the eyes of God, but you will at least say, to use Popper's language, that it is beginning to look as if Meehl's theory has considerable *verisimilitude*, that is, "truth-likeness." (An adequate reconstruction of the verisimilitude concept has yet to be provided by our logician friends, see, e.g., Popper, 1976, but few reflective psychologists will doubt that some such notion of "nearness to the truth" is unavoidable when we evaluate theories. It is crucial to recognize that verisimilitude is an ontological, not an epistemological, concept that must not be conflated with confirmation, probability, evidence, proof, corroboration, belief, support, or plausibility.)

Popperians would speak of low logical or prior probability, of the high content (forbidding much), because it specifies exactly which days it will rain how many inches. A Bayesian (who would reject Popper's philosophy on the grounds that we want our "theoretical prior" to be *high* to get a nice boost out of Bayes' theorem when the facts turn out right) would express Popper's point by saying that we want what Pap (1962, p. 160) calls the *expectedness*, the prior on the observations that is found in the denominator of Bayes' theorem to be low. An unphilosophical chemist or astronomer or molecular biologist would say that this was just good sensible scientific practice, that a theory that makes precise predictions and correctly picks out *narrow intervals* or *point values* out of the range

of experimental possibilities is a pretty strong theory. There are revisions (as I think, necessary) of the classic Popperian position urged on us by his heretical exstudents P. K. Feyerabend and the late Imre Lakatos, but psychologists must reach at least the stage of Bayes and Popper before they can profitably go on to the refinements and criticisms of these gentlemen.

The most important caveat I would adjoin to Sir Karl's falsifiability requirement arises from the considerations pressed by Feyerabend (1962, 1965, 1970, 1971), Lakatos (1970, 1974a, 1974b), and others concerning the crucial role of auxiliary theories in subjecting the main substantive theory of interest to danger of *modus tollens*. As is well-known (and not disputed by Popper), when we spell out in detail the logical structure of what purports to be an observational test of a theoretical conjecture T , we normally find that we cannot get to an observational statement from T alone. We require further a set of often complex and problematic auxiliaries A , plus the empirical realization of certain conditions describing the experimental particulars, commonly labeled collectively as C . So that the derivation of an observation from a substantive theory T amounts always to the longer formula $(T.A.C) \rightarrow O$, rather than the simplified schema $(T \rightarrow O)$ that most of us learned in undergraduate logic courses. This presents a problem not perhaps for Popper's main thesis (although some critics do say this) but for its application as a criterion of the scientific status of theories (or the scientific approach of a particular theoretician or investigator?). The *modus tollens* now reads: Since $(T.A.C) \rightarrow O$, and we have falsified O observationally, we have the consequence $\sim(T.A.C)$. Unfortunately, this result does not entail the falsity of T , the substantive theory of interest but only the falsity of the conjunction $(T.A.C)$; that is, we have proved a disjunction of the falsities of the conjuncts. So the failure to get the expected observation O proves that $\sim T \vee \sim A \vee \sim C$, which is not quite what we would like to show.

One need not subscribe to the famous Duhemian thesis regarding falsification of science as a whole (Grünbaum, 1960, 1962, 1969, 1976) or to the Lakatosian exposition (La-

katos, 1970, 1974a, 1974b) about the protective belt of auxiliaries against which the *modus tollens* is directed versus the hard core of the theory against which the *modus tollens* is, prior to a Kuhnian revolution (Kuhn, 1970a, 1970b, 1970c), forbidden to be directed, to see that there is a difficult problem presented to even a neo-Popperian (like myself), because in social science the auxiliaries *A* and the initial and boundary conditions of the system *C* are frequently as problematic as the theory *T* itself. *Example*: Suppose that a personologist or social psychologist wants to investigate the effect of social fear on visual perception. He attempts to mobilize anxiety in a sample of adolescent males, chosen by their scores on the Social Introversion (*Si*) scale of the Minnesota Multiphasic Personality Inventory (MMPI), by employing a research assistant who is a raving beauty, instructing her to wear Chanel No. 5, and adopt a mixed seductive and castrative manner toward the subjects. An interpretation of a negative empirical result leaves us wondering whether the main substantive theory of interest concerning social fear and visual perception has been falsified, or whether only the auxiliary theories that the *Si* scale is valid for social introversion and that attractive but hostile female experimenters elicit social fear in introverted young males have been falsified. Or perhaps even the particular conditions were not met; that is, she did not consistently act the way she was instructed to or the MMPI protocols were mis-scored.

There is nothing qualitatively unique about this problem for the inexact sciences, but it is quantitatively more severe for us than for the chemist or astronomer, for at least two reasons, which I shall set forth without either proving or developing them here. First, independent testing of the auxiliary theories (which often means validation of psychometric instruments or ascertaining efficacy of social stimulus inputs) is harder to carry out. Due to unavoidable looseness of the nomological network (Cronbach & Meehl, 1955/1973) plus the factors in the list of 20 difficulties *supra*, the range of research circumstances in which auxiliaries *A* are problematic is greater than in the exact sciences or in some but not all of the biological sciences. Second,

a point to which philosophers of science have devoted little attention, in physics or chemistry there is usually a more intimate connection, sometimes one of contributing to derivability, between the substantive theory of interest T and components of the auxiliaries A . This is sometimes even true in advanced branches of biology. *Example*: There is a complicated, well-developed, and highly corroborated theory of how a cyclotron works, and the subject matter of that auxiliary "theory of the instrument" is for the most part identical to the subject matter of the physical theories concerning nuclear particles, and so on, being investigated by the physicist. Devices for bringing about a state of affairs, for isolating the system under study, and for observing what occurs as a result are all themselves legitimated by theory.

It seems there is a sense in which auxiliary theories used by physical and biological scientists are at least subtly informed by what may be loosely called *the spirit*, the leading ideas, the core, and pervasive concepts of the main substantive theory *T*, although not rigorously derivable from *T*. When this is not so, scientists are likely to consider the (*T.A.*) system as "unaesthetic," "incoherent," even *ad hoc*. These fascinating matters remain to be analyzed and reconstructed by logicians, but most scientists and historians of science are—however informally—well aware of their influence. (See, e.g., Holton, 1973.)

In the social sciences, no such intimate connection, and almost never a relation of theoretical derivability, exists; hence, the auxiliary theory (such as a theory that the Rorschach is valid for detecting subclinical schizoid cognitive slippage or that Chanel-doused beauteous research assistants are anxiety elicitors) must stand on its own feet. Almost nothing we know or conjecture about the substantive theory helps us to any appreciable degree in firming up our reliance on the auxiliary. The situation in which *A* is merely conjoined to *T* in setting up our test of *T* makes it hard for us social scientists to fulfill a Popperian falsifiability requirement—to state before the fact what would count as a strong falsifier.

I shall illustrate this problem further with a simple example whose adequate exposition

will appear elsewhere (Golden & Meehl, in press). Suppose that I wish to test my dominant gene conjecture (Golden & Meehl, 1978; Meehl, 1972, 1972/1973g, 1977) concerning *schizotaxia* as the central nervous system condition for the development by social learning of *schizotypy* (Meehl, 1962/1973c), which in turn is the personality precondition for the development of a *clinical schizophrenia*—although the latter must then occur only in one fourth of the persons carrying the gene, given the roughly 12% concordance for first-degree relatives as regards diagnosable clinical schizophrenia. (See also Böök, 1960; Heston, 1966, 1970; Slater, 1958/1971). I might rely on some complex neurological or projective or structured test “sign” as having such-and-such estimated construct validity for the schizotypal personality makeup. Such a quantitative estimate might be made relying on a combination of empirical evidence concerning discordant monozygotic twins of known schizophrenics, protocols of persons tested as college freshmen who subsequently decompensate into a recognizable schizophrenia, and the like. Such numerical estimates will all suffer not only from the usual test unreliability and random sampling fluctuations, but they will also have some unknown degree of systematic bias. For instance, it clearly will not do to assume that the taxon *all compensated schizotypes* would average the same scores on a Rorschach or MMPI indicator variable as do the compensated (discordant) monozygotic twins, the latter being a biased selection, since they have the same potentiating genes that their decompensated twins have. However, there must be something else about them—of an environmental sort—that works strongly in their favor and helps keep them discordant, that is, clinically well. One simply has no way of ascertaining the net impact of these two opposed kinds of forces on the psychometric results.

Suppose that we take some combination of earlier findings on preschizophrenics, remitted schizophrenics, compensated discordant monozygotic twins of schizophrenics, and so forth, and we ascertain that while the valid positive rate p_a among these safely presumed schizotypes varies (even if the sample sizes are huge, it will always vary in an amount unexplainable

by random sampling fluctuation), it nevertheless shows a “reasonably close” agreement. (Again, we think like physicists or physiologists instead of like social scientists fooling around with t tests.) So we strike some kind of rough average \bar{p}_a of these several valid positive rates, knowing that it is the best we can do at this point with data on different groups of schizotypes, who, despite their differences, must all have somehow been tagged as such. Given that estimated valid positive rate, and given a false positive rate p_n (also systematically biased because of the undiagnosed compensated schizotypes in any “control population”), we record our numerical predictions for the incidence of our psychometric sign among parent pairs of schizophrenic probands (where, on the dominant gene theory, we expect not only a 50% schizotypy incidence but something stronger; to wit, at least one member of each parent pair must be a schizotype). We also compute it for siblings and dizygotic twins and—although here things get a bit feeble—with sufficiently large samples, maybe second-degree relatives. Thus, for instance, the expected sign-positive rate among parents (and sibs, if they all cooperate) is given by the simple expression $p^* = \frac{1}{2}p_a + \frac{1}{2}p_n$.

Now the substantive dominant gene theory T , when conjoined with the auxiliary theory A concerning psychometric validity, and assuming that we have identified the right relatives and the probands were all schizophrenics [=C], generates point predictions and therefore takes a high Popperian risk *when the conjunction (T.A.C) is considered as the “theory” under test*. Hence, the verification of those numerical point predictions as to the values of the psychometric incidence in relatives of different degrees of consanguinity provides a strong Popperian test for that conjunctive “theory.” One would then normally say that successful negotiation of this hurdle, the failure to be clobbered *modus tollens* by the outcome of the empirical study, provides a moderate to strong corroboration of the conjunctive theory. Hence, (T.A.C) is doing well; that is, it has escaped falsification despite taking a high risk by making several numerical point predictions.

So far, so good, and Popper as well as his critics would have no complaint. However, the

classical Popperian requirement on playing the scientific game fairly involves the theoretician's saying, before doing the research, what would count as a strong basis for rejecting the theory. If "the theory" is taken to be the substantive theory *T* (which it is, if one is not being philosophically disingenuous) rather than the psychometric auxiliary and diagnostic validity conjectures *A* and *C*, then one will be committing what amounts in spirit to a Popperian sin against falsificationism as a method. If the empirical research does not pan out as predicted, one does not abandon *T*; instead he tells us that either *T* is incorrect, *A* is incorrect, or the diagnoses were untrustworthy!

I am not persuaded from his writings nor from conversations that I have had with him that Sir Karl adequately appreciates the degree to which this theory and auxiliary problem permeate research in the inexact sciences, especially the social sciences in their soft areas. Whether it presents a general problem for the Popperian formulation of scientific method is beyond the scope of this article and my competence. It is perhaps worth saying, however, for the benefit of philosophically oriented readers, that the above described situation—certainly no rarity in our field or in biology—may represent a social fact about the way science works that presents grave difficulties for the Popperian reconstruction. That is, the stipulation beforehand that one will be pleased about substantive theory *T* when the numerical results come out as forecast, but will not necessarily abandon it when they do not, seems on the face of it to be about as blatant a violation of the Popperian commandment as you could commit. For the investigator, in a way, is doing what Popper says we ought not to do, and what astrologers and Marxists and psychoanalysts allegedly do, playing "heads I win, tails you lose." But it seems in accordance with much scientific practice and, as far as I have sampled, with most persons' scientific common sense or intuitions, to say that if the combination (*T.A.C*) generates a high-risk numerical point prediction, such a result really does support all three of the components. The reason it does so seems pretty clear, despite its commonsense, non-formalized character: Because of the lack of intimate inner connection in the inexact sci-

ences between the components of these conjunctions, it would strike us as a *very strange coincidence* if the substantive theory *T* should have low verisimilitude (which would, were *T* true, also generate mispredictions of the numerical point values) and yet the two (largely unrelated) "wrongs" of *T* and *A* are somehow systematically balanced so as to generate the same numerical prediction generated from the conjecture that *T* and *A* both have relatively high verisimilitude.

Such a delicate quantitative counterbalancing of theoretical errors is not impossible, but it seems quite implausible, assuming that nature is (as Einstein says) "subtle but not malicious." So I think we are not being unreasonable to congratulate ourselves on arriving at a successful prediction of high-risk point values or other antecedently improbable observational patterns from the conjunction (*T.A.C*), despite the fact that we seem to be hedging when we say before the fact that we will not consider our substantive theory *T* falsified by a bad result if it does not pan out. These are problems that need further exploration by statisticians and philosophers of science, especially in light of work on the history of science, and with special attention to the question of whether there are important differences between the inexact and the exact sciences, or even between the biological and social sciences, as to how a Popperian or neo-Popperian methodology should be explained and applied.

But, you may say, what has all this got to do with significance testing? Isn't the social scientist's use of the null hypothesis simply the application of Popperian (or Bayesian) thinking in contexts in which probability plays such a big role? No, it is not. One reason it is not is that the usual use of null hypothesis testing in soft psychology as a means of "corroborating" substantive theories does not subject the theory to grave risk of refutation *modus tollens*, but only to a rather feeble danger. The kinds of theories and the kinds of theoretical risks to which we put them in soft psychology when we use significance testing as our method are *not* like testing Meehl's theory of weather by seeing how well it forecasts the number of inches it will rain on certain days. Instead, they are depressingly close

to testing the theory by seeing whether it rains in April at all, or rains several days in April, or rains in April more than in May. It happens mainly because, as I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false. I shall not attempt to document this here, because among sophisticated persons it is taken for granted. (See Morrison & Henkel, 1970, especially the chapters by Bakan, Hogben, Lykken, Meehl, and Rozeboom.) A little reflection shows us why it has to be the case, since an output variable such as adult IQ, or academic achievement, or effectiveness at communication, or whatever, will always, in the social sciences, be a function of a sizable but finite number of factors. (The smallest contributions may be considered as essentially a random variance term.) In order for two groups (males and females, or whites and blacks, or manic depressives and schizophrenics, or Republicans and Democrats) to be *exactly* equal on such an output variable, we have to imagine that they are exactly equal *or* delicately counterbalanced on all of the contributors in the causal equation, which will never be the case.

Following the general line of reasoning (presented by myself and several others over the last decade), from the fact that the null hypothesis is always false in soft psychology, it follows that the probability of refuting it depends wholly on the sensitivity of the experiment—its logical design, the net (attenuated) construct validity of the measures, and, most importantly, the sample size, which determines where we are on the statistical power function. Putting it crudely, if you have enough cases and your measures are not totally unreliable, the null hypothesis will always be falsified, *regardless of the truth of the substantive theory*. Of course, it could be falsified in the wrong direction, which means that as the power improves, the probability of a corroborative result approaches one-half. However, if the theory has no verisimilitude—such that we can imagine, so to speak, picking our empirical results randomly out of a directional hat apart from any theory—the probability of refuting by getting a significant difference in the wrong direction also approaches one-half. Obviously, this is quite un-

like the situation desired from either a Bayesian, a Popperian, or a commonsense scientific standpoint. As I have pointed out elsewhere (Meehl, 1967/1970b; but see criticism by Oakes, 1975; Keuth, 1973; and rebuttal by Swoyer & Monson, 1975), an improvement in instrumentation or other sources of experimental accuracy tends, in physics or astronomy or chemistry or genetics, to subject the theory to a greater risk of refutation *modus tollens*, whereas improved precision in null hypothesis testing usually decreases this risk. A successful significance test of a substantive theory in soft psychology provides a feeble corroboration of the theory because the procedure has subjected the theory to a feeble risk.

But, you may say, we do not look at just one; we look at a batch of them. Yes, we do; and how do we usually do it? In the typical *Psychological Bulletin* article reviewing research on some theory, we see a table showing with asterisks (hence, my title) whether this or that experimenter found a difference in the expected direction at the .05 (one asterisk), .01 (two asterisks!), or .001 (three asterisks!!) levels of significance. Typically, of course, some of them come out favorable and some of them come out unfavorable. What does the reviewer usually do? He goes through what is from the standpoint of the logician an almost meaningless exercise; to wit, he *counts noses*. If, say, Fisbee's theory of the mind has a batting average of 7:3 on 10 significance tests in the table, he concludes that Fisbee's theory seems to be rather well supported, "although further research is needed to explain the discrepancies." This is scientifically a preposterous way to reason. It completely neglects the crucial asymmetry between confirmation, which involves an inference in the formally invalid third figure of the implicative syllogism (this is why inductive inferences are ampliative and dangerous and why we can be objectively wrong even though we proceed correctly), and refutation, which is in the valid fourth figure, and which gives the *modus tollens* its privileged position in inductive inference. Thus the adverse tests, seen properly, do Fisbee's theory far more damage than the favorable ones do it good.

I am not making some nit-picking statistician's correction. I am saying that the whole business is so radically defective as to be scientifically almost pointless. This is not a technical hassle about whether Fisbee should have used the varimax rotation, or how he estimated the communalities, or that perhaps some of the higher order interactions that are marginally significant should have been lumped together as a part of the error term, or that the covariance matrices were not quite homogeneous. I am not a statistician, and I am not making a statistical complaint. I am making a philosophical complaint or, if you prefer, a complaint in the domain of scientific method. I suggest that when a reviewer tries to "make theoretical sense" out of such a table of favorable and adverse significance test results, what the reviewer is actually engaged in, willy-nilly or unwittingly, is meaningless substantive constructions on the properties of the statistical power function, and almost nothing else.

This feckless activity is made worse by the almost universal practice of what I call *step-wise low validation*. By this I mean that we rely on one investigation to "validate" a particular instrument and some other study to validate another instrument, and then we correlate the two instruments and claim to have validated the substantive theory. I do not argue that this is a scientific nothing, but it is about as close to a nothing as you can get without intending to. Consider that I first show that Meehl's Mental Measure has a validity coefficient (against the criterion I shall here for simplicity take to be quasi-infallible or definitive) of, say, .40—somewhat higher than we usually get in personology and social psychology! Then I show that Glotz's Global Gauge has a validity for its alleged variable of the same amount. Relying on these results, having stated the coefficient and gleefully recorded the asterisks showing that these coefficients are not zero (!), I now try to corroborate the Glotz-Meehl theory of personality by showing that the two instruments, each having been duly "validated," correlate .40, providing, happily, some more asterisks in the table. Now just what kind of a business is this? Let us suppose that each instrument has a reliability of .90 to make it easy. That means that

the portion of construct-valid variance for each of the devices is around one fifth of the reliable variance and the same for their overlap when correlated with each other. I do not want to push the discredited (although recently revived) principle of indifference, but without other knowledge, it is easily possible, and one could perhaps say rather likely, that the correlation between the two occurs in a region of each one's components that has literally nothing to do with either of the two criterion variables used in the validity studies relied on. This is, of course, especially dangerous in light of the research that we have on the contribution of methods variance.

I seem to have trouble conveying to my students and colleagues just how dreadful a mess of flabby inferences this kind of thing involves. It is as if we were interested in the effect of sunlight on the mating behavior of birds, but not being able to get directly at either of these two things, we settle for correlating a proxy variable like field-mice density (because the birds tend to destroy the field mice) with, say, incidence of human skin cancer (since you can get that by spending too much time in the sun!) You may think this analogy dreadfully unfair; but I think it is a good one. Of course, the whole idea of simply counting noses is wrong, because a theory that has seven facts for it and three facts against it is *not* in good shape, and it would not be considered so in any developed science.

You may say, "But, Meehl, R. A. Fisher was a genius, and we all know how valuable his stuff has been in agronomy. Why shouldn't it work for soft psychology?" Well, I am not intimidated by Fisher's genius, because my complaint is not in the field of mathematical statistics; and as regards inductive logic and philosophy of science, it is well-known that Sir Ronald permitted himself a great deal of dogmatism. I remember my amazement when the late Rudolf Carnap said to me, the first time I met him, "But, of course, on this subject Fisher is just mistaken; surely you must know that." My statistician friends tell me that it is not clear just how useful the significance test has been in biological science either, but I set that aside as beyond my competence to discuss. The shortest answer to this rebuttal about agronomy, and one that has general im-

portance in thinking about soft psychology, is that we must carefully distinguish *substantive theory* from *statistical hypothesis*. There is a tendency in the social sciences to conflate these in talking about our inferences. (A neglected article by Bolles, 1962, did not cure the psychologists' disease.) The substantive theory is the theory about the causal structure of the world, the entities and processes underlying the phenomena; the statistical hypothesis is a much more restricted and "operational" conjecture about the value of some parameter, such as the mean of a specified statistical population. The main point in agronomy is that the logical distance, the difference in meaning or content, so to say, between the alternative hypothesis and substantive theory *T* is so small that only a logician would be concerned to distinguish them. *Example*: I want to find out whether I should be putting potash on the ground to help me raise more corn. Now everybody knows from common sense as well as biology that the corn gets its nutrients from the soil, and furthermore that the yield of corn at harvest time is not causally efficacious in determining what I did in the spring, random numbers aside. If I refute the statistical null hypothesis that plots of corn with potash do not differ in yield from plots without potash, I have thereby proved the alternative hypothesis—that there *is* a difference between these two sorts of plots; and the only substantive conclusion to draw, given such a difference, is that the potash made the difference. Such a situation, in which the content of the substantive theory is logically quasi-identical with the alternative hypothesis, which was refuted by our significance test, is completely different from the situation in soft psychology. Fisbee's substantive theory of the mind is not equivalent, or anywhere near equivalent, to the alternative hypothesis. All sorts of competing theories are around, including my grandmother's common sense, to explain the nonnull statistical difference. So the psychologist can take little reassurance about the use of significance tests from knowing that Fisher's approach has been useful in studying the effect of fertilizer on crop yields.

Although this presents a pretty depressing picture, I daresay that the Skinner disciples among you will be inclined to think,

well, that's just one more way of showing what we have known all along. The point is to prove that you have achieved experimental control over your subject matter, as Skinner says. If you have, I am not much interested in tabular asterisks; if you haven't, I'm not interested in them either.

But that is easy for Skinnerians because their theory (it is a theory in Sir Karl Popper's sense) is close to a pure dispositional theory, and does not usually present us with the kind of evidentiary evaluation problem that we get with entity-postulating theories such as those of Freud, Hull, Albert Ellis, or, to come closer to home, my conjectures about schizophrenia or hedonic deficit (Meehl, 1972, 1974, 1975, 1962/1973c, 1972/1973g). Those of us whose cognitive passions are incompletely satisfied by dispositional theories, whether Skinnerian or psychometric, should ask ourselves what *kind* of inferred entity construction we want and how it could generate the sorts of intellectual "surprises" that Robert Nozick (1974, pp. 18–22) considers typical of invisible hand theories, which have proved so eminently successful in the physical and biological sciences and—somewhat less so—in economics. Some directions of solution (before I go on to the one that I am using in my own research) follow.

We could take the complex form of Bayes's theorem more seriously in concrete application to various substantive theories to take into account, even if crudely in the sense of setting upper and lower bounds to the probabilities involved, the logical asymmetry between confirmation and refutation (see, e.g., Maxwell, 1974). Second, it may be that the Fisherian tradition, with its soothing illusion of quantitative rigor, has inhibited our search for stronger tests, so we have thrown in the sponge and abandoned hope of concocting substantive theories that will generate stronger consequences than merely "the Xs differ from the Ys." Thus, for instance, even when we cannot generate numerical point predictions, (the ideal case found in the exact sciences), it may be that we can at least predict the order of numerical values or the rank order of the first-order numerical differences, and the like.

Sometimes in the other sciences it has been possible to concoct a middling weak theory that, while incapable of generating numerical

point values, entails a certain *function form*, such as a graph should be an ogive or that it should have three peaks and that these peaks should be increasingly high, and that the distance on the abscissa between the first two peaks should be less than the distance between the second two. In the early history of quantum theory, physicists relied on Wien's law, which related "some (unknown) function" of wavelength to energy multiplied by the fifth power of wavelength. In the cavity radiation experiment, the empirical points were simply plotted at varying temperatures, and it was evident by inspection that they fell on the same curve, even though a formal expression for that curve was beyond the theory's capabilities (Eisberg, 1961, pp. 50-51).

Talking of Wien's law is a good time for me to recommend to psychologists who disagree with my position to have a look at any textbook of theoretical chemistry or physics, where one searches in vain for a statistical significance test (and finds few confidence intervals). The power of the physicist does not come from exact assessment of probabilities that a difference exists (which physicists would view as a ludicrous thing to show), nor by the verbal precision of so-called "operational definitions" in the embedding text. The physicist's scientific power comes from two other sources, namely, the immense deductive fertility of the formalism and the accuracy of the measuring instruments. The scientific trick lies in conjoining rich mathematics and experimental precision, a sort of "invisible hand wielding fine calipers." The embedding text is sometimes surprisingly loose, free-wheeling, even metaphorical—as viewers of television's *Nova* are aware, seeing Nobel laureates discourse whimsically about the charm, strangeness, and gluons of nuclear particles (see, e.g., Nambu, 1976). One gets the impression that when you have a good science going, with potent mathematics and accurate instruments, you can be relaxed and easygoing about the words. Nothing is as stuffy and pretentious as the verbal "pseudorigor" of the soft branches of social science. In my modern physics text, I am unable to find one single test of statistical significance. What happens instead is that the physicist has a sufficiently powerful invisible hand theory that enables him to generate an

expected curve for his experimental results. He plots the observed points, looks at the agreement, and comments that "the results are in reasonably good accord with theory." Moral: *It is always more valuable to show approximate agreement of observations with a theoretically predicted numerical point value, rank order, or function form, than it is to compute a "precise probability" that something merely differs from something else.* Of course, we do not have precise probabilities when we do significance testing because of the falsity of the assumptions generating the table's values and varying robustness of our tests under departures from these assumptions.

The only possible "solution" to the theory-refutation problem that I have time to discuss in any detail is what I call *consistency tests* (Meehl, Note 3). Unfortunately, this approach is not easily available for most theoretical problems in soft psychology, although I am not prepared to say that it is confined to the domain in which I have been developing it, namely, taxometrics, that is, the application of psychometric procedures to depiction of a taxonic situation and classification of individuals into the taxon or outside of it. From our conjectures about the latent causal situation, we derive formulas for estimating the theoretical quantities of interest, such as the proportion of schizotypes in a given clinical population, the mean values of the schizotypal and nonschizotypal classes, the optimal cut ("hitmax") on each phenotypic indicator variable for classifying individuals, and the proportion of valid and false positives achieved by that cut. But we realize that our conjectures about the latent situation may be false or that the indicators relied on may have too low validity, or that they may be more correlated within the taxa than desired, and so forth. Second, even if the basic formal structure postulated is approximated by the state of nature (e.g., there is a schizoid taxon, the indicators have sizable validity, the intra-taxon distributions are quasi-normal or at least unimodal, the correlation of the indicators within the groups is small, and the departures from these various hypotheses are within the tolerance allowed by the method's robustness), it may still be that we have suffered some kind of systematic bias on one of

the indicators due to a nuisance variable such as social class, or that we have had bad luck in the sample, so the method's numerical deliverances on this occasion are untrustworthy.

Whether the abstract causal structure postulated is unsound or the numerical values found in this sample are seriously in error, we need some method of checking the data internally to find out whether these unfortunate possibilities have materialized. We do this by deriving theorems within the formalism specifying how various numerical values (observed or calculated from the observed) should be related to each other, so that when they are not related as the consistency theorem demands, we are alerted to the danger that something is rotten in the state of Denmark (see Meehl, 1973d). Unfortunately, most of the work, both mathematical and empirical, is as yet only available in mimeographed reports from our laboratory (Golden, 1976; Golden & Meehl, Note 1, Note 2; Meehl, Note 3, Note 4). What survives scrutiny will be found in a book in preparation with my former student and research colleague Robert Golden (Golden & Meehl, in press).

One taxometric procedure, which I have christened *maxcov-hitmax* (Meehl, 1973d) relies on the following theorem: If three fallible indicator variables are negligibly correlated within a diagnostic taxon and within the extra taxon population, then the covariance of any pair of these is maximized in that class interval on the third indicator that contains the hitmax (optimal, fewest misses) cut on the third indicator. That is, $\text{cov}(yz)$ has its largest value for the subset of patients falling in the hitmax interval on x . Starting from this relation we go through a sequence of calculations yielding estimates of the base rate P of the taxon, the frequency distributions of all three of the fallible indicators, the location of all three hitmax cuts, and the inverse probability of taxon membership (via Bayes' theorem) for a patient who has any given combination of the three signs plus or minus.

Our Monte Carlo runs and our single application to a real case in which we know the true answer and pretend not to know it, namely, biological sex diagnosed by three MMPI femininity keys, have been most encouraging and suggest that the method is

powerful and quite robust under departures from the simplifying hypotheses. But applying it to a situation in which we do not know the true answer (such as "What is the proportion of unrecognized schizotypes in a mixed psychiatric population?"), how much faith should we have in our numerical results? The best way I know to go about this, since mere replication of the inferred parameter estimates does not answer the question, is by the use of consistency tests. For example, one of the consistency tests in this kind of two-category taxonic situation is this: If we form the product of the differences between the inferred latent means on y and z (schizotypes minus non-schizotypes) and then multiply this product $\Delta\bar{y}\Delta\bar{z}$ by the product of the inferred schizotypal base-rate P and its complement Q , then it can be shown that this theoretically calculated quantity should equal the grand covariance of y and z computed directly from the observations. We call this the "total covariance consistency test."

Of course, such a relation is not required to be literally true, because it is known in advance that (a) the impoverished theory has imperfect verisimilitude and (b) all statistical estimates are subject to both systematic and random error. (We are *not* going to do a significance test!) What we have is a problem of robustness and detection of excessive departures from the postulated latent conditions. Golden and I arbitrarily said that we would consider a particular sample as delivering sufficiently accurate information if the estimates of base rate and hit rate were within .10 of the true values, and estimated latent means and standard deviations within one class interval of the truth. (Actually we did much better than that on the average. For example, with sample sizes greater than 400, equal variances, two sigma differences of latent means, and zero intrataxon correlations, the average error for P was only .01 and for latent means and sigmas, less than one fourth standard deviation which is one-half the smallest integral class interval.) But if these tolerances strike you as excessively large, I remind you how much more powerful such numerical claims are in soft psychology than the usual flabby "the boys are taller than the girls" or "the schizophrenics are shyer than the manic

Table 1
Description of Sample Sets

Description of Sample Sets											
Set	Variable	<i>N</i>	<i>P</i>	<i>M</i> _o	<i>M</i> _t	<i>SD</i> _o	<i>SD</i> _t	<i>D'</i>	<i>SD</i> _t / <i>SD</i> _o	<i>r</i>	<i>F</i> ^a
1.1	<i>N</i>	1,000	.5	8	12	2	2	2	1	0 ^b	0
1.2		800	.5	8	12	2	2	2	1	0 ^b	0
1.3		600	.5	8	12	2	2	2	1	0 ^b	0
1.4		400	.5	8	12	2	2	2	1	0 ^b	0
2.1	<i>P</i>	1,000	.6	8	12	2	2	2	1	0 ^b	3
2.2		1,000	.7	8	12	2	2	2	1	0 ^b	2
2.3		1,000	.8	8	12	2	2	2	1	0 ^b	8
2.4		1,000	.9	8	12	2	2	2	1	0	0
3.1	<i>D'</i>	1,000	.5	9	12	2	2	1.5	1	0 ^b	0
3.2		1,000	.5	10	12	2	2	1	1	0 ^b	15
3.3		1,000	.5	11	12	2	2	.5	1	0	0
3.4		1,000	.5	12	12	2	2	0	1	0	0
4.1	<i>SD</i> _t / <i>SD</i> _o	1,000	.5	8	12	1.9	2.1	2	1.1	0 ^b	0
4.2		1,000	.5	8	12	1.7	2.3	2	1.3	0 ^b	0
4.3		1,000	.5	8	12	1.5	2.5	2	1.7	0 ^b	0
4.4		1,000	.5	8	12	1	3	2	3	0	0
5.1	<i>r</i>	1,000	.5	8	12	2	2	2	1	.1 ^b	0
5.2		1,000	.5	8	12	2	2	2	1	.3 ^b	0
5.3		1,000	.5	8	12	2	2	2	1	.5 ^b	8
5.4		1,000	.5	8	12	2	2	2	1	.8	0
<i>r</i> _o / <i>r</i> _t											
6.1	<i>r</i> _o / <i>r</i> _t = 4 <i>N</i>	1,000	.8	8	12	2	2	2	1	.5/.125	0
6.2		800	.8	8	12	2	2	2	1	.5/.125	0
6.3		600	.8	8	12	2	2	2	1	.5/.125	0
6.4		400	.8	8	12	2	2	2	1	.5/.125	0

Note. N = sample size; P = base rate of the taxon; M_o = mean of the extra taxon class on each indicator; M_t = mean of the taxon on each indicator; SD_o = standard deviation of the extra taxon class on each indicator; SD_t = standard deviation of the taxon on each indicator; $D' = (M_t - M_o)/S$, where $S = SD_o + SD_t/2$; r = latent correlation between indicator pairs; F = number of failures of consistency tests in 25 samples.

^a 94% correct.

^b Parameter estimates are always or nearly always accurate.

depressives." We then imposed tolerances on each of the four most promising consistency tests derived within the formalism. For example, if the total covariance consistency test $T_1 = \text{cov}(yz) - PQ (\bar{y}_s - \bar{y}_n) (\bar{z}_s - \bar{z}_n)$ yields a discrepancy greater than $.64 + .74s^2$, a "robustness cut" chosen by a combination of analytical derivation with preliminary Monte Carlo trials, then this particular sample is considered "numerically inconsistent" with Consistency Test T_1 . Now if any one of the four consistency tests is, so to speak, rejected by a given sample, this is a red flag warning

us that we ought not to have much faith in the parametric estimates of interest.

The important question then is, how sensitive are the consistency tests to sample departures from the parametric truth in excess of the tolerance allowed? How often will we draw a sample in which the inferred parameters are in error by more than the tolerance limit imposed but all four consistency tests are satisfied within their tolerance limits, leading us mistakenly to trust our results? Second, how often is at least one of the four consistency tests numerically inconsistent (i.e.,

Table 2
Consistency Test Result

Actual situation	Sample		Total
	Trust-worthy	Suspicious	
Accurate	336	36	372
Inaccurate	0	228	228
Total	336	264	600

outside its tolerance limit) leading us to *mis-trust* the sample when in fact all of the sample estimates of the parameters are within their tolerances? The first of these we might call a "false negative" failure on the part of the consistency tests to function jointly; the second is then a false positive.

I restrict my data presentation to Monte Carlo runs in which the samples are generated from a multivariate normal model, although I want to emphasize that our methods are not generally confined to the normal case. Normality was imposed because of Monte Carlo generating problems. In Table 1, the numbers "Set 1.1, 1.2, . . ." in the first column merely name conditions of fixed population properties and sample sizes, and 25 Monte Carlo samples were drawn per set. The column heads indicate the various population properties, such as taxon base-rate P , the two latent taxon means and standard deviations, the mean difference in standard deviation units, the ratio of latent standard deviations, and the within-group correlations. The important result (F) indicates how many of the 25 samples under the given set conditions were failures of the consistency tests. Thus, the four consistency tests were applied to each sample, which was classified as probably trustworthy (or probably not) in accordance with the tolerance rules for consistency tests. Then the sample was classified as to whether it was *in fact* trustworthy, that is, whether the main latent parameters were all estimated within their allowed tolerance.

Despite the high average accuracy of our taxometric method when evaluated as mean percent errors in estimating each of the latent parameters (base rate, hit rates, means, standard deviations), if a naive trusting taxometrist relied blindly on the method, hoping to

be accurate on all seven parameters on any sample drawn, he would be misled distressingly often were he to lack consistency tests. Among our 600 Monte Carlo samples, all seven latent parameters of the artificial population were estimated to an accuracy within the tolerance levels in 372 samples; that is, on 228 samples at least one parameter was inaccurate. This shows that a trustworthy device for detecting such bad samples is much to be desired. It will not do a taxonomic scientist much good to be "usually quite accurate" if the procedure relied on is nevertheless often (38% of the time) somewhat inaccurate *and the investigator is without a method that warns him when the untoward event has, on a given occasion, occurred.*

In Table 2 the 600 Monte Carlo samples are tallied with respect to each sample's parameter estimation accuracy and whether it passed all four consistency tests. It is encouraging that overall the consistency tests were 94% accurate. Furthermore, the 6% of the samples in which the consistency tests erred were all samples in which they erred conservatively; that is, one or more of the consistency tests was suspiciously outside its tolerance limits, yet none of the latent parameters estimated by the methods was outside *its* tolerance limits. We have not as yet drawn a single Monte Carlo sample (among 600) in which the four consistency tests were conjunctively reassuring but the sample was in fact misleading. This finding suggests that we were unduly stringent, so that if some small amount of leeway were permitted for errors of the other kind, the consistency tests could be somewhat relaxed and, perhaps concurrently, the tolerance limits on the parameter estimates could be somewhat tightened.

There is some interchangeability between original estimators and consistency tests, and the maxcov-hitmax method itself was originally derived by me as a consistency test before I realized that it could better be used as an original search device (see Meehl, Note 3, pp. 28-29; Note 4, pp. 2-6).

Not in reliance on these results, which I present merely as exemplars of a general methodological thesis, I want now to state as strongly as I can a prescription that we should adopt in soft psychology to help get away

from the feeble practice of significance testing: *Wherever possible, two or more non-redundant estimates of the same theoretical quantity should be made, because multiple approximations to a theoretical number are always more valuable, provided that methods of setting permissible tolerances exist, than a so-called exact test of significance, or even an exact setting of confidence intervals.* This is a special case of what my philosopher colleague Herbert Feigl refers to as "triangulation in logical space." It is, as you know, standard procedure in the developed sciences. We have, for instance, something like a dozen independent ways of estimating Avogadro's number, and since they all come out "reasonably close" (again, I have never seen a physicist do a t test on such a thing!), we are confident that we know how many molecules there are in a mole of chlorine.

This last point may lead you to ask, "If consistency tests are as important as Meehl makes them out to be, why we don't hear about them in chemistry and physics?" I have a perfect answer to that query. It goes like this: *Consistency tests are so much a part of standard scientific method in the developed disciplines, taken so much for granted by everybody who researches in chemistry or physics or astronomy or molecular biology or genetics, that these scientists do not even bother having a special name for them!* It shows the sad state of soft psychology when a fellow like me has to cook up a special meta-theory expression to call attention to something that in respectable science is taken as a matter of course.

Having presented what seems to me some encouraging data, I must nevertheless close with a melancholy reflection. The possibility of deriving consistency tests in the taxonic situation rests on the substantive problems presented by fields like medicine and behavior genetics, and it is not obvious how we would go about doing this in soft areas that are non-taxonic. It may be that the nature of the subject matter in most of personology and social psychology is inherently incapable of permitting theories with sufficient conceptual power (especially mathematical development) to yield the kinds of strong refuters expected by Popperians, Bayesians, and unphilosophical

scientists in developed fields like chemistry. This might mean that we could most profitably confine ourselves to low-order inductions, a (to me, depressing) conjecture that is somewhat corroborated by the fact that the two most powerful forms of clinical psychology are atheoretical psychometrics of prediction on the one hand and behavior modification on the other. Neither of these approaches has the kind of conceptual richness that attracts the theory-oriented mind, but I think we ought to acknowledge the possibility that there is never going to be a really impressive theory in personality or social psychology. I dislike to think that, but it might just be true.

Addendum

My colleague, Thomas J. Bouchard, Jr., on reading a draft of this article faulted me for what he saw as a major inconsistency between my neo-Popperian emphasis on falsifiability and my positive assessment of Freud. There is no denying that for such a quantitatively oriented product of the "dust-bowl empiricist" tradition as myself, I do have a soft spot in my heart (Minnesota colleagues would probably say in my head) for psychoanalysis. So, the most honest and straightforward way to deal with Bouchard's complaint might be simply to admit that the evidence on Freud is inadequate and that Bouchard and I are simply betting on different horses. But I cannot resist the impulse to say just a bit more on this vexatious question, because while I am acutely aware of a pronounced (and possibly irrational) difference in the "educated prior" I put on Freud as contrasted with rubber band theory or labeling theory or whatever, I am not persuaded that my position is as grossly incoherent as it admittedly appears. Passing the question whether attempts to study psychoanalytic theory by the methods of experimental or differential psychology have on the whole tended to support rather than refute it (see, e.g., Fisher & Greenberg, 1977; Rapa-port, 1959; Sears, 1943; Silverman, 1976), my own view is that the best place to study psychoanalysis is the psychoanalytic session itself, as I have elsewhere argued in a far too condensed way (Meehl, 1970/1973e).

I believe that some aspects of psychoana-

lytic theory are not presently researchable because the intermediate technology required—which really means instruments-cum-theory—does not exist. I mean auxiliaries and methods such as a souped-up, highly developed science of psycholinguistics, and the kind of mathematics that is needed to conduct a rigorous but clinically sensitive and psychoanalytically realistic job of theme tracing in the analytic protocol. This may strike some as a kind of cop-out, but I remind you that Lakatos, Kuhn, Feyerabend, and others have convincingly made the point that there are theories in the physical and biological sciences that are untestable when first propounded because the theoretical and technological development necessary for making certain kinds of observations bearing on them had not taken place. It is vulgar positivism (still held by many psychologists) to insist that any respectable empirical theory must be testable, if testable means *definitively testable right now*.

But I do think that there is another class of consequences of psychoanalytic theory, close to the original "clinical connections" alleged by Freud, Ferenczi, Jones, Abraham, and others that does not involve much of what Freud called *the witch metapsychology*, where no complicated statistics are needed, let alone the invention of any *new* formal modes of protocol analysis. Here the problem is mainly that *none of us has bothered to carry out some relatively simple-minded kinds of analyses on a random sample of psychoanalytic protocols collected from essentially naive patients to whom no interpretations have as yet been offered*. This second category is, in my view, a category of research studies that we could have done, but have not done. *Example*: We can easily ascertain whether manifest dream content of a certain kind is statistically associated (in the simple straightforward sense of a patterned fourfold table) with such and such kinds of thematic material in the patient's subsequent associations to the dream. I would not even object to doing significance tests on a batch of such tables, but to explain why would unduly enlarge what is already an addendum.

I cheerfully admit, in this matter, to the presence of a large distance between my subjective personalistic probability (based on my

experiences as analysand and practitioner of psychoanalytic therapy) and the present state of the "intersubjective public evidence." That is what I mean by saying that Bouchard and I are betting on different horses. But one must distinguish, as I know from subsequent conversations that he does, between a criticism (a) that what *is* proper evidence *does* presently exist and is *adverse* to a conjecture and (b) an anti-Popperian claim that falsifiability in principle does not matter. If I thought (as does Popper) that Freudian theory was in principle not falsifiable, then I would have to confess to a major inconsistency. But I do think it is falsifiable, although I agree that *some parts* of it cannot *at present* be tested because of the primitive development of the auxiliary theories and the measurement technologies that would be jointly necessary.

A final point on this subject is one that I hesitate to include because it is very difficult to explain in the present state of philosophy of science, and I could be doing my main thesis damage by presenting a cursory and somewhat dogmatic statement of it. Nevertheless, having made the above statements about psychoanalytic theory and having contrasted it favorably with some of the (to me, trivial and flabby) theories in soft psychology, I fear I have an obligation to say it, however ineptly. Once one sees that it is inappropriate to conflate the concepts *rational* and *statistical*, then it is a fuzzy open question, in the present state of the metatheoretician's art, just when a mass of nonquantitative converging evidence can be said to have made a stronger case for a conjecture than the weak kinds of nonconverging quantitative evidence usually represented by the significance testing tradition. I say "when" rather than "whether," because it is blindingly obvious that *sometimes* qualitative evidence of certain sorts is superior in its empirical weight to what a typical social, personality, or clinical psychologist gets in support of a substantive theory by the mere refutation of the null hypothesis. Take, for instance, the evidence in a well-constructed criminal case, such as the evidence that Bruno Hauptmann was the kidnapper of the Lindbergh baby. I do not see how anybody who reads the trial transcript of the Hauptmann

case could have a reasonable doubt that he was guilty as charged. Yet I cannot recall any of the mass of data that convicted him as being of a quantitative sort (one cannot fairly except the serial numbers on the gold notes, they being not "measures" but "football numbers").

All of us believe a lot of things that we would not have the vaguest idea how to express as a probability value (*pace* strong Bayesians!) or how to compute as an indirect test of statistical significance. I believe, for instance, that Adolf Hitler was a schizotype; I do not believe that Kaspar Hauser was the son of a prince; I believe that the domestic cat probably was evolved from *Felis lybica* by the ancient Egyptians; I hold that my sainted namesake wrote the letter to the Corinthians but did not write the letter to the Hebrews; I am confident that my wife is faithful to me; and so forth. The point is really a simple one—that there are many areas of both practical and theoretical inference in which nobody knows how to calculate a numerical probability value, and nobody knows how to state the manner or degree in which various lines of evidence converge on a certain conjecture as having high verisimilitude. There are propositions in history (such as, "Julius Caesar crossed the Rubicon") that we all agree are well corroborated by the available documents but without any *t* tests or the possibility of calculating any, whereas Fisbee's theory of social behavior is only weakly corroborated by the fact that he got a significant *t* test when he compared the boys and the girls or the older kids and the younger kids on the Hockheimer-Sedlitz Communication Scale. Now I consider my betting on the horse of psychoanalysis to be in the same kind of ball park as my beliefs about Julius Caesar or the evolution of the cat. But, I repeat, this may be a terribly irrational leap of faith on my part. For the purposes of the present article and Bouchard's criticism of it, I hope it is sufficient to say that one could arguably hold that significance testing in soft psychology is a pretentious endeavor that falls under a tolerant neo-Popperian criticism, and could nevertheless enter his personalistic prediction that *when adequate tests become available to us, a sizable portion of psychoanalytic theory*

will escape refutation. So I do not think I am actually contradicting myself, but I am personalistically betting on the outcome of a future horse race.

Reference Notes

1. Golden, R., & Meehl, P. E. *Detecting latent clinical taxa, IV: Empirical study of the maximum covariance method and the normal minimum chi-square method, using three MMPI keys to identify the sexes* (Tech. Rep. PR-73-2). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1973. (a)
2. Golden, R., & Meehl, P. E. *Detecting latent clinical taxa, V: A Monte Carlo study of the maximum covariance method and associated consistency tests* (Tech. Rep. PR-73-3). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1973. (b)
3. Meehl, P. E. *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion* (Tech. Rep. PR-65-2). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1965.
4. Meehl, P. E. *Detecting latent clinical taxa, II: A simplified procedure, some additional hitmax cut locators, a single-indicator method, and miscellaneous theorems* (Tech. Rep. PR-68-4). Minneapolis: University of Minnesota Psychiatry Research Laboratory, 1968.

References

- Allport, G. W. *Personality: A psychological interpretation*. New York: Holt, 1937.
- Andreski, S. *Social sciences as sorcery*. London: Deutsch, 1972.
- Barker, R. G. *Ecological psychology*. Stanford, Calif.: Stanford University Press, 1968.
- Bartlett, M. S. *An introduction to stochastic processes*. Cambridge, England: Cambridge University Press, 1955.
- Bolles, R. C. The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, 1962, 11, 639-645.
- Böök, J. A. Genetical aspects of schizophrenic psychoses. In D. D. Jackson (Ed.), *The etiology of schizophrenia*. New York: Basic Books, 1960.
- Braithwaite, R. B. *Scientific explanation*. New York: Harper, 1960.
- Braun, J. R. (Ed.). *Clinical psychology in transition* (Rev. ed.). Cleveland, Ohio: World, 1966.
- Brunswick, E. Representative design and probabilistic theory. *Psychological Review*, 1955, 62, 236-242.
- Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Campbell, N. R. *Foundations of science*. New York: Dover Publications, 1957. (Originally published as *Physics, the elements*, 1920.)

- Carnap, R. Testability and meaning. New Haven, Conn.: Yale University Graduate Philosophy Club, 1950. (Originally published, 1936-1937.)
- Carnap, R. Meaning postulates. In R. Carnap, *Meaning and necessity*. Chicago: University of Chicago Press, 1956. (Originally published, 1952.)
- Carnap, R. *Philosophical foundations of physics*. New York: Basic Books, 1966.
- Cartwright, D. Determinants of scientific progress: The case of the risky shift. *American Psychologist*, 1973, 28, 222-231.
- Cronbach, L. J. The two disciplines of scientific psychology. *American Psychologist*, 1957, 12, 671-684.
- Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press, 1973. (Originally published, 1955.)
- Eisberg, R. M. *Fundamentals of modern physics*. New York: Wiley, 1961.
- Feigl, H. Some major issues and developments in the philosophy of science of logical empiricism. In H. Feigl & M. Scriven (Eds.), *The foundations of science and the concepts of psychology and psychoanalysis: Minnesota studies in the philosophy of science* (Vol. 1). Minneapolis: University of Minnesota Press, 1956.
- Feller, W. *An introduction to probability theory and its applications* (2nd ed.). New York: Wiley, 1957.
- Feyerabend, P. K. Explanation, reduction, and empiricism. In H. Feigl & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science* (Vol. 3): *Scientific explanation, space and time*. Minneapolis: University of Minnesota Press, 1962.
- Feyerabend, P. K. Problems of Empiricism, Part I. In R. G. Colodny (Ed.), *Beyond the edge of certainty*. Englewood Cliffs, N.J.: Prentice-Hall, 1965.
- Feyerabend, P. K. Against method: Outline of an anarchistic theory of knowledge. In M. Radner & Winokur (Eds.), *Minnesota studies in the philosophy of science* (Vol. 4): *Analyses of theories and methods of physics and psychology*. Minneapolis: University of Minnesota Press, 1970.
- Feyerabend, P. K. Problems of empiricism, Part II. In R. G. Colodny (Ed.), *The nature and function of scientific theories*. Pittsburgh, Pa.: University of Pittsburgh Press, 1971.
- Fisher, F. M. *The identification problem in econometrics*. New York: McGraw-Hill, 1966.
- Fisher, R. A. *Statistical methods and scientific inference*. Edinburgh, Scotland: Oliver & Boyd, 1956.
- Fisher, R. A. *The design of experiments* (8th ed.). Edinburgh, Scotland: Oliver & Boyd, 1966.
- Fisher, R. A. *Statistical methods for research workers* (13th ed.). Edinburgh, Scotland: Oliver & Boyd, 1967.
- Fisher, S., & Greenberg, P. *The scientific credibility of Freud's theories and therapy*. New York: Basic Books, 1977.
- Fiske, D. W. The limits of the conventional science of personality. *Journal of Personality*, 1974, 24, 1-11.
- Gergen, K. J. Social psychology as history. *Journal of Personality and Social Psychology*, 1973, 26, 309-320.
- Golden, R. *Psychometric verisimilitude*. Unpublished doctoral dissertation, University of Minnesota, 1976.
- Golden, R. R., & Meehl, P. E. Testing a single dominant gene theory without an accepted criterion variable. *Annals of Human Genetics*, 1978, in press.
- Golden, R. R., & Meehl, P. E. *Taxometric analysis of causal entities: Detection of the schizoid taxon*. New York: Academic Press, in press.
- Gottesman, I. I. Heritability of personality: A demonstration. *Psychological Monographs*, 1963, 77(9, Whole No. 572).
- Gottesman, I. I., & Shields, J. *Schizophrenia and genetics: A twin study vantage point*. New York: Academic Press, 1972.
- Grünbaum, A. The Duhemian argument. *Philosophy of Science*, 1960, 11, 75-87.
- Grünbaum, A. Falsifiability of theories: Total or partial? *Synthese*, 1962, 14, 17-34.
- Grünbaum, A. Can we ascertain the falsity of a scientific hypothesis? *Studium Generale*, 1969, 22, 1061-1093.
- Grünbaum, A. Ad hoc auxiliary hypotheses and falsificationism. *British Journal for the Philosophy of Science*, 1976, 27, 329-362.
- Gunther, M. *The luck factor*. New York: Macmillan, 1977.
- Harrison, T. R., et al. (Eds.), *Principles of internal medicine*. New York: McGraw-Hill, 1966.
- Hempel, C. G. *Fundamentals of concept formation in empirical science*. Chicago: University of Chicago Press, 1952.
- Hempel, C. G. The theoretician's dilemma. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science* (Vol. 2): *Concepts, theories, and the mind-body problem*. Minneapolis: University of Minnesota Press, 1958.
- Heston, L. L. Psychiatric disorders in foster home reared children of schizophrenic mothers. *British Journal of Psychiatry*, 1966, 112, 819-825.
- Heston, L. L. The genetics of schizophrenia and schizoid disease. *Science*, 1970, 167, 249-256.
- Hinde, R. A. *Animal behavior* (2nd ed.). New York: McGraw-Hill, 1970.
- Hogan, R., DeSoto, C. B., & Solano, C. Traits, texts, and personality research. *American Psychologist*, 1977, 32, 255-264.
- Holt, R. R. Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology*, 1958, 56, 1-12.
- Holton, G. *Thematic origins of scientific thought: Kepler to Einstein*. Cambridge, Mass.: Harvard University Press, 1973.
- Jencks, C. *Inequality*. New York: Basic Books, 1972.
- Johnson, W. E. *Logic, Part I*. New York: Dover Publications, 1964. (Originally published, 1921.)
- Kemeny, J. G., Snell, J. L., & Thompson, G. L. *Introduction to finite mathematics*. Englewood Cliffs, N.J.: Prentice-Hall, 1957.
- Keuth, H. On prior probabilities of rejecting statisti-

- cal hypotheses. *Philosophy of Science*, 1973, 40, 538-546.
- Kordig, R. The comparability of scientific theories. *Philosophy of Science*, 1971, 38, 467-485.
- Kordig, C. R. Discussion: Observational invariance. *Philosophy of Science*, 1973, 40, 558-569.
- Kuhn, T. S. Logic of discovery or psychology of research? In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, England: Cambridge University Press, 1970. (a)
- Kuhn, T. S. Reflections on my critics. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, England: Cambridge University Press, 1970. (b)
- Kuhn, T. S. *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press, 1970. (c)
- Lakatos, I. Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, England: Cambridge University Press, 1970.
- Lakatos, I. Popper on demarcation and induction. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper* (Vol. 1). LaSalle, Ill.: Open Court, 1974. (a)
- Lakatos, I. The role of crucial experiments in science. *Studies in History and Philosophy of Science*, 1974, 4, 309-325. (b)
- Langmuir, I. Science, common sense and decency. *Science*, 1943, 97, 1-7.
- Lazarus, A. A. *Behavior therapy and beyond*. New York: McGraw-Hill, 1971.
- Lewis, D. How to define theoretical terms. *Journal of Philosophy*, 1970, 67, 427-446.
- Li, C. C. *Path analysis*. Pacific Grove, Calif.: Boxwood Press, 1975.
- Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports Monograph*, 1957, 9, 635-694.
- London, I. D. Some consequences for history and psychology of Langmuir's concept of convergence and divergence of phenomena. *Psychological Review*, 1946, 53, 170-188.
- MacCorquodale, K., & Meehl, P. E. Edward C. Tolman. In W. K. Estes et al., *Modern learning theory: A critical analysis of five examples*. New York: Appleton-Century-Crofts, 1954.
- Maxwell, G. Meaning postulates in scientific theories. In H. Feigl & G. Maxwell (Eds.), *Current issues in the philosophy of science*. New York: Holt, Rinehart & Winston, 1961.
- Maxwell, G. The necessary and the contingent. In H. Feigl & G. Maxwell (Eds.), *Scientific explanation, space and time: Minnesota studies in the philosophy of science* (Vol. 3). Minneapolis: University of Minnesota Press, 1962.
- Maxwell, G. Corroboration without demarcation. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper*. LaSalle, Ill.: Open Court, 1974.
- McGuire, W. J. The yin and yang of progress in social psychology: Seven koans. *Journal of Personality and Social Psychology*, 1973, 26, 446-456.
- Meehl, P. E. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press, 1954.
- Meehl, P. E. Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science* (Vol. 4) *Analyses of theories and methods of physics and psychology*. Minneapolis: University of Minnesota Press, 1970. (a)
- Meehl, P. E. Theory-testing in psychology and physics: A methodological paradox. In D. E. Morrison & R. E. Henkel (Eds.), *The significance test controversy*. Chicago: Aldine, 1970. (b) (Originally published, 1967.)
- Meehl, P. E. A critical afterword. In I. I. Gottesman & J. Shields, *Schizophrenia and genetics*. New York: Academic Press, 1972.
- Meehl, P. E. The cognitive activity of the clinician. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press, 1973. (a) (Originally published, 1960.)
- Meehl, P. E. High school yearbooks: A reply to Schwarz. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press, 1973. (b) (Originally published, 1971.)
- Meehl, P. E. Schizotaxia, schizotypy, schizophrenia. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press, 1973. (c) (Originally published, 1962.)
- Meehl, P. E. MAXCOV-HITMAX: A taxonomic search method for loose genetic syndromes. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press, 1973. (d)
- Meehl, P. E. Some methodological reflections on the difficulties of psychoanalytic research. *Psychological Issues*, 1973, 8, 104-115. (e) (Originally published, 1970.)
- Meehl, P. E. Some ruminations on the validation of clinical procedures. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press, 1973. (f) (Originally published, 1959.)
- Meehl, P. E. Specific genetic etiology, psychodynamics, and therapeutic nihilism. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press, 1973. (g) (Originally published, 1972.)
- Meehl, P. E. Why I do not attend case conferences. In P. E. Meehl, *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press, 1973. (h)
- Meehl, P. E. Genes and the unchangeable core. *Voices: The Art and Science of Psychotherapy*, 1974, 38, 25-35.
- Meehl, P. E. Hedonic capacity: Some conjectures. *Bulletin of the Menninger Clinic*, 1975, 39, 295-307.
- Meehl, P. E. Specific etiology and other forms of strong influence: Some quantitative meanings. *Journal of Medicine and Philosophy*, 1977, 2, 33-53.
- Mischel, W. On the future of personality measurement. *American Psychologist*, 1977, 32, 246-254.

- Morrison, D. E. & Henkel, R. E. (Eds.). *The significance test controversy*. Chicago: Aldine, 1970.
- Nagel, F. *The structure of science*. New York: Harcourt, Brace & World, 1961.
- Nambu, Y. The confinement of quarks. *Scientific American*, 1976, 235, 48-60.
- Nozick, R. *Anarchy, state and utopia*. New York: Basic Books, 1974.
- Oakes, W. F. On the alleged falsity of the null hypothesis. *Psychological Record*, 1975, 25, 265-272.
- Pap, A. Reduction-sentences and open concepts. *Methodos*, 1953, 5, 3-30.
- Pap, A. *Semantics and necessary truth*. New Haven, Conn.: Yale University Press, 1958.
- Pap, A. *An introduction to the philosophy of science*. New York: Free Press, 1962.
- Popper, K. R. *The logic of scientific discovery*. New York: Basic Books, 1959.
- Popper, K. R. *Conjectures and refutations*. New York: Basic Books, 1962.
- Popper, K. R. *Objective knowledge*. Oxford, England: Oxford University Press, 1972.
- Popper, K. R. *Autobiography*. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper*. LaSalle, Ill.: Open Court, 1974.
- Popper, K. R. A note on verisimilitude. *British Journal for the Philosophy of Science*, 1976, 27, 147-195.
- Ramsey, F. P. Theories. In R. B. Braithwaite (Ed.), *F. P. Ramsey's The foundations of mathematics and other logical essays*. Paterson, N.J.: Littlefield, Adams, 1960. (Originally published, 1931.)
- Rapaport, D. The structure of psychoanalytic theory: A systematizing attempt. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3). *Formulations of the person and the social context*. New York: McGraw-Hill, 1959.
- Read, R. C. *A mathematical background for economists and social scientists*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.
- Schilpp, P. A. (Ed.), *The philosophy of Karl Popper*. LaSalle, Ill.: Open Court, 1974.
- Schlenker, B. R. Social psychology and science. *Journal of Personality and Social Psychology*, 1974, 29, 1-15.
- Sears, R. R. *Survey of objective studies of psychoanalytic concepts*. New York: *Social Science Research Council Bulletin* No. 51, 1943.
- Sellars, W. Concepts as involving laws and inconceivable without them. *Philosophy of Science*, 1948, 15, 287-315.
- Sells, S. B. An interactionist looks at the environment. *American Psychologist*, 1963, 18, 696-702.
- Silverman, L. H. Psychoanalytic theory: "The reports of my death are greatly exaggerated." *American Psychologist*, 1976, 31, 621-637.
- Skinner, B. F. *The behavior of organisms*. New York: Appleton-Century, 1938.
- Slater, E. The monogenic theory of schizophrenia. In J. Shields & I. I. Gottesman (Eds.), *Man, mind, and heredity: Selected papers of Eliot Slater on psychiatry and genetics*. Baltimore, Md.: Johns Hopkins University Press, 1971. (Originally published, 1958.)
- Smith, M. B. Criticisms of a social science. *Science*, 1973, 180, 610-612.
- Stoddard, L. *Luck, your silent partner*. New York: Liveright, 1929.
- Swoyer, C., & Monson, T. C. Theory confirmation in psychology. *Philosophy of Science*, 1975, 42, 487-502.
- Vico, G. *The new science of Giambattista Vico* (3rd rev. ed.) (T. C. Bergin and M. H. Fisch, trans.). Ithaca: Cornell University Press, 1948. (Originally published, 1744.)
- Waismann, F. Verifiability. *Proceedings of the Aristotelian Society*, 1945, 19, 119-150.
- Wiggins, J. Despair and optimism in Minneapolis. *Contemporary Psychology*, 1973, 18, 605-606.

Received December 12, 1977 ■

A Reader's, Writer's, and Reviewer's Guide to Assessing Research Reports in Clinical Psychology

Brendan A. Maher
Harvard University

The Editors of the *Journal of Consulting and Clinical Psychology* who served between 1974 and 1978 have seen some 3,500 manuscripts in the area of consulting and clinical psychology. Working with this number of manuscripts has made it possible to formulate a set of general guidelines that may be helpful in the assessment of research reports. Originally developed by and for journal reviewers, the guidelines are necessarily skeletal and summary and omit many methodological concerns. They do, however, address the methodological concerns that have proved to be significant in a substantial number of cases. In response to a number of requests, the guidelines are being made available here.

Topic Content

1. Is the article appropriate to this journal? Does it fall within the boundaries mandated in the masthead description?

Style

1. Does the manuscript conform to APA style in its major aspects?

Introduction

1. Is the introduction as brief as possible given the topic of the article?
2. Are all of the citations correct and necessary, or is there padding? Are important citations missing? Has the author been careful to cite prior reports contrary to the current hypothesis?
3. Is there an explicit hypothesis?
4. Has the *origin* of the hypothesis been made explicit?
5. Was the hypothesis *correctly* derived from the theory that has been cited? Are other, contrary hypotheses compatible with the same theory?
6. Is there an explicit rationale for the selection of measures, and was it derived logically from the hypothesis?

Method

1. Is the method so described that replication is possible without further information?
2. **Subjects:** Were they sampled randomly from the population to which the results will be generalized?
3. Under what circumstances was informed consent obtained?
4. Are there probable biases in sampling (e.g., volunteers, high refusal rates, institution population atypical for the country at large, etc.)?
5. What was the "set" given to subjects? Was there deception? Was there control for experimenter influence and expectancy effects?
6. How were subjects debriefed?
7. Were subjects (patients) led to believe that they were receiving "treatment"?
8. Were there special variables affecting the subjects, such as medication, fatigue, and threat that were not part of the experimental manipulation? In clinical samples, was "organicity" measured and/or eliminated?
9. **Controls:** Were there appropriate control groups? What was being controlled for?
10. When more than one measure was used, was the order counterbalanced? If so, were order effects actually analyzed statistically?
11. Was there a control task(s) to confirm specificity of results?
12. **Measures:** For both dependent and independent variable measures—was validity and reliability established and reported? When a measure is tailor-made for a study, this is very important. When validities and reliabilities are already available in the literature, it is less important.
13. Is there adequate description of tasks, materials, apparatus, and so forth?
14. Is there discriminant validity of the measures?
15. Are distributions of scores on measures typical of scores that have been reported for similar samples in previous literature?
16. Are measures free from biases such as
 - a. Social desirability?
 - b. Yeasaying and naysaying?
 - c. Correlations with general responsivity?
 - d. Verbal ability, intelligence?
17. If measures are scored by observers using categories or codes, what is the inter-rater reliability?
18. Was administration and scoring of the measures done blind?
19. If short versions, foreign-language translations, and so forth, of common measures are used, has the validity and reliability of these been established?
20. In correlational designs, do the two measures have theoretical and/or methodological independence?

Representative Design

1. When the stimulus is a human (e.g., in clinical judgments of clients of differing race, sex, etc.), is there a *sample* of stimuli (e.g., more than one client of each race or each sex)?
2. When only one stimulus or a few human stimuli were used, was an adequate explanation of the failure to sample given?

Statistics

1. Were the statistics used with appropriate assumptions fulfilled by the data (e.g., normalcy of distributions for parametric techniques)? Where necessary, have scores been transformed appropriately?
2. Were tests of significance properly used and reported? For example, did the author use the p value of a correlation to justify conclusions when the actual size of the correlation suggests little common variance between two measures?
3. Have statistical significance levels been accompanied by an analysis of practical significance levels?
4. Has the author considered the effects of a limited range of scores, and so forth, in using correlations?
5. Is the basic statistical strategy that of a "fishing expedition"; that is, if many comparisons are made, were the obtained significance levels predicted in advance? Consider the number of significance levels as a function of the total number of comparisons made.

Factor Analytic Statistics

1. Have the correlation and factor matrices been made available to the reviewers and to the readers through the National Auxiliary Publications Service or other methods?
2. Is it stated what was used for communalities and is the choice appropriate? Ones in the diagonals are especially undesirable when items are correlated as the variables.
3. Is the method of termination of factor extraction stated, and is it appropriate in this case?
4. Is the method of factor rotation stated, and is it appropriate in this case?
5. If items are used as variables, what are the proportions of yes and no responses for each variable?
6. Is the sample size given, and is it adequate?
7. Are there evidences of distortion in the final solution, such as singlet factors, excessively high communalities, obliqueness when an orthogonal solution is used, linearly dependent variables, or too many complex variables?
8. Are artificial factors evident because of inclusion of variables in the analysis that are alternate forms of each other?

Figures and Tables

1. Are the figures and tables (a) necessary and (b) self-explanatory? Large tables of nonsignificant differences, for example, should be eliminated if the few obtained significances can be reported in a sentence or two in the text. Could several tables be combined into a smaller number?
2. Are the axes of figures identified clearly?
3. Do graphs correspond logically to the textual argument of the article? (E.g., if the text states that a certain technique leads to an *increment* of mental health and the accompanying graph shows a *decline* in symptoms, the point is not as clear to the reader as it would be if the text or the graph were amended to achieve visual and verbal congruence.)

Discussion and Conclusion

1. Is the discussion properly confined to the findings or is it digressive, including new post hoc speculations?
2. Has the author explicitly considered and discussed viable alternative explanations of the findings?
3. Have nonsignificant trends in the data been promoted to "findings"?
4. Are the limits of the generalizations possible from the data made clear? Has the author identified his/her own methodological difficulties in the study?
5. Has the author "accepted" the null hypothesis?
6. Has the author considered the possible methodological bases for discrepancies between the results reported and other findings in the literature?

Many detailed responses to a first draft were reviewed. Particular acknowledgment is due to Thomas Achenbach, George Chartier, Andrew Comrey, Jesse Harris, Mary B. Harris, Alan Kazdin, Richard Lanyon, Eric Mash, Martha Mednick, Peter Nathan, K. Daniel O'Leary, N. D. Reppucci, Robert Rosenthal, Richard Suinn, and Norman Watt.

Requests for reprints should be sent to Brendan A. Maher, Department of Psychology and Social Relations, Harvard University, Cambridge, Massachusetts 02138.

This material may be reproduced in whole or in part without permission, provided that acknowledgment is made to Brendan A. Maher and the American Psychological Association.

CHARLES C THOMAS • PUBLISHER

NEXUS PSYCHOTHERAPY: Between Humanism and Behaviorism by Kenneth Urial Gutsch, *Univ. of Southern Mississippi, Hattiesburg*, and Jacob Virgil Ritenour, II, *Jackson, Michigan*. Foreword by Howard S. Rosenblatt and Harold H. Dawley, Jr. '78, 196 pp., 15 il., 16 tables, \$14.75

CEREBRAL LOGIC: Solving the Problem of Mind and Brain by Charles W. Needham, *Neurological Associates of Tucson, Tucson, Arizona*. '78, 232 pp., 1 il., \$12.75

DEVELOPMENTAL DISABILITIES OF EARLY CHILDHOOD edited by Barbara A. Feingold and Caryl L. Bank, both of *Little Village School, Merrick, New York*. (19 Contributors) '78, 216 pp., 14 il., 4 tables, \$16.50

BASIC APPROACHES TO GROUP PSYCHOTHERAPY AND GROUP COUNSELING (2nd Ed., 2nd Ptg.) edited by George M. Gazda, *Univ. of Georgia, Athens*. (20 Contributors) '77, 560 pp., 9 il., \$16.75

HYPNOSIS INDUCTION TECHNICS (3rd Ptg.) by Myron Teitelbaum. Foreword by Michael M. Gilbert. '77, 200 pp., \$8.50

UP FROM UNDERACHIEVEMENT by Gary S. Felton and Barbara E. Biggs, both of the *Center for Interpersonal Studies, Los Angeles*. '77, 208 pp., 13 il., 9 tables, cloth-\$10.50, paper-\$6.95

PRESCRIPTIONS FOR CHILDREN WITH LEARNING AND ADJUSTMENT PROBLEMS (5th Ptg.) by Ralph F. Blanco, *Temple Univ., Philadelphia, Pennsylvania*. '77, 320 pp., \$10.50

Principles and Techniques of INTERVENTION WITH HYPERACTIVE CHILDREN edited by Marvin J. Fine, *Univ. of Kansas, Lawrence*. (11 Contributors) '77, 328 pp., 44 il., \$21.50

THE PARAPROFESSIONAL IN THE TREATMENT OF ALCOHOLISM: A New Profession (2nd Ptg.) edited by George E. Staub, *Alcohol Abuse and Alcoholism Program, Los Angeles*, and Leona M. Kent, *San Joaquin County Alcoholism Program, Stockton, California*. (12 Contributors) '76, 184 pp., 1 table, \$8.50

SPEECH AND LANGUAGE DELAY: A Home Training Program (3rd Ed.) by R. Ray Battin, *The Battin Clinic*; C. Olaf Haug, *Private Practice*; and Cristine Blair and Susan Miller, both of *The Battin Clinic*; all of *Houston, Texas*. Foreword by W. Leonard Draper. '78, 120 pp., 24 il., 1 table, \$10.50

COUNSELING PARENTS OF THE MENTALLY RETARDED: A Sourcebook (3rd Ptg.) compiled and edited by Robert L. Noland. (39 Contributors) '78, 420 pp., 4 il., 18 tables, \$11.75

DIRECT OBSERVATION AND MEASUREMENT OF BEHAVIOR (3rd Ptg.) by S. J. Hutt and Corrine Hutt, both of *Park Hospital for Children, Oxford, England*. Foreword by I. Newton Kugelmass. '78, 240 pp. (6 3/4 x 9 3/4), 81 il., 14 tables, \$13.50

GROUP COUNSELING AND GROUP PSYCHOTHERAPY WITH REHABILITATION CLIENTS edited by Milton Seligman, *Univ. of Pittsburgh, Pittsburgh, Pennsylvania*. (18 Contributors) '77, 352 pp., 1 il., 7 tables, cloth-\$16.50, paper-\$12.75

REFLEX EPILEPSY, BEHAVIORAL THERAPY AND CONDITIONAL REFLEXES by Francis M. Forster, *Univ. of Wisconsin, Madison*. '77, 328 pp., 87 il., 22 tables, \$22.75

THE HYPERACTIVE CHILD IN THE CLASSROOM by Frank P. Alabiso, *Milton Robinson and Associates, Springville, New York*, and James C. Hansen, *State Univ. of New York, Buffalo*. '77, 336 pp., 10 il., \$15.75

PSYCHOTHERAPY FOR WOMEN: Treatment Toward Equality edited by Edna I. Rawlings, *Univ. of Cincinnati, Cincinnati, Ohio*, and Dianne K. Carter, *Univ. of Iowa, Iowa City*. (22 Contributors) '77, 500 pp. (6 3/4 x 9 3/4), 10 il., 16 tables, \$17.50

THE EMOTIONALLY TROUBLED CHILD: A Guide for Parents and Teachers in the Early Recognition of Mental and Nervous Disorders in Children by Robert L. Mason, Jr., Bert O. Richmond and Lucien B. Fleurant, all of *Univ. of Georgia, Athens*. '76, 196 pp., \$11.50

—Orders with remittance sent, on approval, postpaid—
301-327 East Lawrence Avenue • Springfield • Illinois • 62717



Association for Advancement of Behavior Therapy

TWELFTH ANNUAL CONVENTION
CONRAD HILTON **CHICAGO, ILLINOIS**
November 17-19, 1978

PRE-CONVENTION INSTITUTES - NOVEMBER 16, 1978
*Intensive 5-hour sessions with leaders in the field, limited
to thirty participants to insure interaction*

**WORKSHOPS - SYMPOSIA - INVITED ADDRESSES
POSTER SESSIONS**

Topics will include:

Treatment of Sexual Dysfunction
Treatment of Addictive Behaviors
Assertion and Social Skills Training
Behavioral Medicine - Treatment of Depression
Self-Control and Cognitive Behavior Modification
Behavior Modification in Business and Industry
Behavioral Approaches to Marital and
Interpersonal Problems

8-HOUR BASIC COURSE IN BEHAVIOR THERAPY
EMPLOYMENT PLACEMENT BUREAU
FILM PROGRAM - EXHIBIT AREA

SOCIETY OF BEHAVIORAL MEDICINE
organizational meeting

Continuing Education Credit including Category I AMA credit for the
Physicians' Recognition Award (17 hours) will be available.

*For a copy of the Convention Program Booklet when it is
published in September please contact:*

Gene G. Abel, M.D., Program Chairperson
AABT, 420 Lexington Avenue, New York, NY 10017

Psychosocial Treatment of Chronic Mental Patients

Milieu vs. Social-Learning Programs

Gordon L. Paul and Robert J. Lentz

Paul and Lentz have spent seven years evaluating alternative methods for treating chronic patients. Here is the definitive report of their startling results.

"There has not been previously available to the field such a matchless rigor in experimental design, so clear a specification of treatment programs, so lengthy a period of in-hospital observation and community after-care follow-up, such precise measurement of dependent variables, and such striking treatment results." —Norman Garnezy, University of Minnesota \$22.50

Harvard University Press

Cambridge, Mass. 02138

Career Opportunities for Psychologists

Expanding and Emerging Areas
Paul J. Woods



May 1976. 26 chapters, 320 pages.
\$5. ISBN 0-912704-03-9.

For the undergraduate and graduate student in psychology.

For the psychologist thinking about changing careers or expanding his/her areas of specialization.

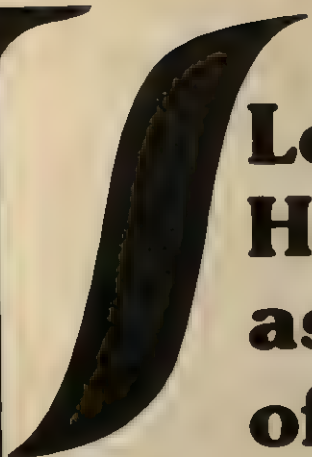
This book reviews the current employment picture for psychologists in academia; explores innovative career opportunities in such areas of employment as health research, population and environment, industry, forensic psychology, engineering psychology, criminal justice, drug and alcohol abuse, public affairs psychology, program evaluation, architecture, and more; includes "how to" chapters on using job placement mechanisms, writing a vita, and writing letters to prospective employers; provides brief job descriptions on psychologists already serving in innovative or nontraditional roles.

Ordering information:

Please send full payment to:



Order Department
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036
(202) 833-7600



Learned Helplessness as a Model of Depression


**Special issue of the
Journal of Abnormal Psychology,
February 1978.**

The issue serves as a forum to investigate the validity and adequacy of the learned helplessness model of depression in humans and includes both critical and supportive theoretical and empirical articles with comments.

L. Rowell Huesmann is guest editor. Richard A. Depue, Martin E. P. Seligman, Charles G. Costello, Robert C. Smolen, and Alexander M. Buchwald are among the contributors.

Copies of the special issue are available at \$6 each, prepaid. Discounts on bulk orders are also available. To order single copies or for more information on bulk orders write:

**Subscription Department
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036**



Evaluation of Psychological Service Delivery Programs



Special Issue of Professional Psychology

Evaluation of Psychological Service Delivery Programs is the title of a special November, 1977 issue of Professional Psychology. Editor Allan Barclay and guest Co-editors Bob and Evelyn Perloff have assembled a distinguished collection of articles that brings readers up to date on the evaluation of psychological services. The issue addresses the need to justify the effectiveness of service delivery programs not only in terms of cost but more importantly in terms of human benefits.

Sample Topics and Contributors

A sample of the topics and contributors appearing in this issue includes: Theodore Blau on quality of life; Nicholas Cummings on psychotherapy; Howard Blane on alcoholism; S. B. Sells and associates on drug abuse, and NIMH's Bertram Brown, Benjamin Liptzin and James Stockdill on the federal view of mental health program evaluation.

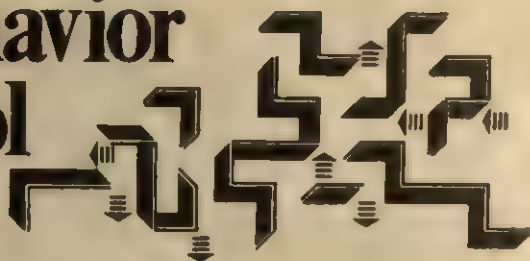
Ordering Information

Single copies of the issue are available from the Order Department at APA for \$5 each. All orders \$25 or under must be prepaid. Please make checks payable to APA. Send your orders to:



Order Department, P.P.
American Psychological
Association
1200 17th Street, N.W.
Washington, D.C. 20036

Master Lectures on Behavior Control



A continuing education program for behavioral scientists. Eight 2-hour lectures presented by eminent psychologists who discuss advances, problems, major concepts, methodologies, and findings in the area of behavior control.

All lectures are in tape cassette format. They may be purchased for \$10 each or the entire 8 lecture set is available at \$75. Payment or an institutional purchase order must accompany all orders.

Other lectures in the series include *Master Lectures on Physiological Psychology*, and *Master Lectures on Developmental Psychology*. Write for a free brochure.



Master Lecture Series/ BC
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036



Irwin Altman, Environmental Control of Behavior and Behavioral Control of the Environment.

Elliot S. Valenstein, The Brain and Behavior Control: Scientific and Social Issues.

William J. McGuire, Persuasion and Social Control.

R. Kirk Schwitzgebel, Legal Issues in Behavior Control.

Nathan H. Azrin, Behavior Modification.

Theodore X. Barber, Hypnosis and Behavior Control.

Daniel X. Freedman, Pharmacology and Behavior Control.

Nicholas Hobbs, Moral and Ethical Issues in Behavior Control.

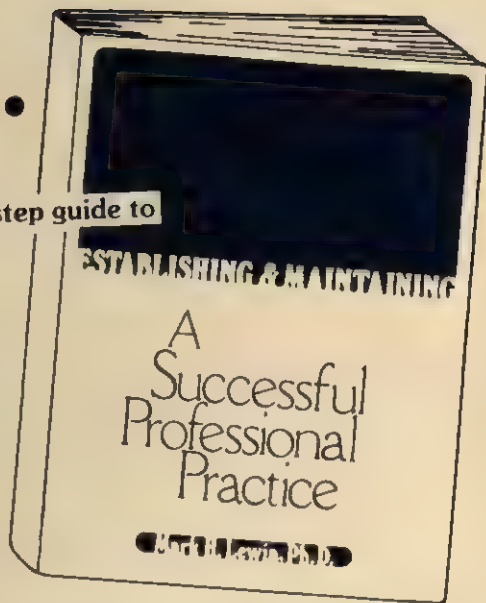
For the First Time...

An easy reading, practical, step by step guide to

A must for

- Full and Part-time private practitioners.
- Those who have ever even considered entering private practice.

"Reduces the chance of failure or disappointment."



An entertaining, well organized book complete with worksheets, plans, and practical approaches to the everyday "nuts and bolts" business aspects of private practice.

MARK H. LEWIN, an ABPP Diplomate, distinguished author and editor, and founder of a thriving independent psychological service center, gives step-by-step procedures for:

- Assessing your readiness for private practice
- Preplanning the practice
- Setting up the office
- Developing new referrals
- Managing your time
- Effective use of lawyers and accountants
- Evaluating the practice as it develops
- Budgeting, fee setting, controlling costs and tax expenses

First Printing 1978 - \$25.00
10% Discount Available Now
\$22.50

Mail checks payable to:
Professional Development Institute
756 East Main Street
Rochester, N.Y. 14605

Please send me _____ copy(ies) of **Establishing and Maintaining a Successful Professional Practice** for which I enclose \$22.50 per book plus New York sales tax where applicable.

Name _____

Street Address _____

City _____ State _____ Zip _____

7/12
78

Could you use this well-known test for brain injury and perceptual-motor coordination, widely employed in clinical practice and in research since 1945, with norms based on large clinical and control groups?

Graham-Kendall

MEMORY FOR DESIGNS TEST

Then order Tester's Set consisting of durable plates and manual (\$17.00).

•

Obtain further information by writing

**PSYCHOLOGICAL TEST
SPECIALISTS**

Box 9229

Missoula, Montana 59807

VOLUME 46, NUMBER 5

OCTOBER 1978

*120
books
in*

*185 me
Library 97 m/2/77*

JOURNAL OF
CONSULTING
AND
CLINICAL
PSYCHOLOGY

AMERICAN PSYCHOLOGICAL ASSOCIATION
ISSN 0022-006X

Editor

Brendan A. Maher, *Harvard University*

Associate Editor

Judith P. Worell, *University of Kentucky*

Editorial Assistant: Zoe Forbes

Consulting Editors

Thomas M. Achenbach, *National Institute of Mental Health, Bethesda, Maryland*

Kent G. Bailey, *Virginia Commonwealth University*

Barbara B. Brown, *Beverly Hills, California*

James N. Butcher, *University of Minnesota*

Jean Chapman, *University of Wisconsin—Madison*

George M. Chartier, *Arizona State University*

Andrew L. Comrey, *University of California, Los Angeles*

W. Grant Dahlstrom, *University of North Carolina, Chapel Hill*

Anthony Davids, *Brown University*

Gerald C. Davison, *State University of New York at Stony Brook*

Alan S. DeWolfe, *Loyola University of Chicago*

Barbara Dohrenwend, *Graduate School and University Center of the City University of New York*

Richard M. Eiser, *Virginia Polytechnic Institute and State University, Blacksburg, Virginia*

Norman S. Endler, *York University, Downsview, Canada*

Sol Garfield, *Washington University in St. Louis*

Donna M. Gelfand, *University of Utah*

Goldine C. Gleser, *University of Cincinnati Medical Center*

Harrison G. Gough, *University of California, Berkeley*

Jeane G. Harris, Jr., *University of Kentucky*

Mary B. Harris, *University of New Mexico*

Kenneth Heller, *Indiana University*

Frances A. Hill, *University of Montana*

Alexandra G. Kaplan, *University of Massachusetts—Amherst*

Alan S. Kaufman, *University of Georgia*

Alan E. Kazdin, *Pennsylvania State University*

Jeffrey A. Kelly, *University of Mississippi Medical Center*

Barbara K. Keogh, *University of California, Los Angeles*

James C. Kincannon, *University of Minnesota*

Richard I. Lanyon, *Arizona State University*

Herbert M. LeCourt, *University of Waterloo, Waterloo, Canada*

Peter M. Lewinsohn, *University of Oregon*

Michael J. Mahoney, *Pennsylvania State University*

G. Alan Marlatt, *University of Washington*

Albert Marston, *University of Southern California*

Eric J. Mash, *University of Calgary, Calgary, Canada*

Richard McFall, *University of Wisconsin—Madison*

Martha T. S. Mednick, *University of Connecticut*

Barbara G. Melamed, *University of Florida*

David I. Mostofsky, *Boston University*

Benjamin J. Murawski, *Peter Bent Brigham Hospital, Boston, Massachusetts*

Peter E. Nathan, *Rutgers—The State University*

John M. Neale, *State University of New York at Stony Brook*

K. Daniel O'Leary, *State University of New York at Stony Brook*

Oscar A. Parsons, *University of Oklahoma*

Nolan E. Penn, *University of California, San Diego*

N. Dickon Reppucci, *University of Virginia*

Jerome Resnick, *Temple University*

Leonard G. Rorer, *Miami University*

Gerald M. Rosen, *Providence Family Medical Center, Seattle, Washington*

Robert Rosenthal, *Harvard University*

William Schofield, *University Hospitals, Minneapolis, Minnesota*

Carolyn Simmons, *University of Colorado at Denver*

Bonnie Strickland, *University of Massachusetts—Amherst*

Hans H. Strupp, *Vanderbilt University*

Richard M. Suinn, *Colorado State University*

Richard S. Swit, *Duke University Medical Center*

Patricia B. Sutker, *Medical University of South Carolina*

Norman F. Watt, *University of Massachusetts—Amherst*

Robert L. Weiss, *University of Oregon*

Joan Welkowitz, *New York University*

G. Terence Wilson, *Rutgers—The State University*

Patricia Wisocki, *University of Massachusetts—Amherst*

Melvin Zax, *University of Rochester*

APA Journal Staff

Anita DeVivo, *Executive Editor*

Ann I. Mahoney, *Manager, Journal Production*

Barbara R. Richman, *Production Editor*

Robert J. Hayward, *Advertising Representative*

Juanita Brodie, *Subscription Manager*

The *Journal of Consulting and Clinical Psychology* publishes original contributions on the following topics: (a) the development, validity, and use of techniques of diagnosis and treatment in disordered behavior; (b) studies of populations of clinical interest, such as hospital, prison, rehabilitation, geriatric, and similar samples; (c) cross-cultural and demographic studies of interest for the behavior disorders; (d) studies of reliability and of its assessment and development where these have a clear bearing on problems of consulting and clinical psychology; or (e) case studies pertinent to the preceding topics. In order to improve the use of journal resources it has been agreed by the two Editors concerned that the *Journal of Consulting and Clinical Psychology* will no longer consider manuscripts dealing with the etiology or descriptive pathology of abnormal behavior, and the *Journal of Abnormal Psychology* will no longer consider articles dealing with the diagnosis or treatment of abnormal behavior. Articles that appear to have a significant contribution to both of these broad areas may be sent to either journal for editorial decision. Papers of a theoretical nature will also be considered within the space limitations of the journal.

Manuscripts: Submit manuscripts in triplicate to the Editor-elect, Sol L. Garfield, Department of Psychology, Washington University, St. Louis, Missouri 63130 according to instructions provided elsewhere in the journal. Authors submitting a manuscript previously considered for publication in another APA journal are invited to inform the Editor, who will then seek to obtain independent reviews, thus avoiding the possibility of repeated reviewing by the same consultant.

Copyright and Permission: All rights reserved. Written permission must be obtained from the American Psychological Association for copying or reprinting text of more than 500 words, tables, or figures. Permissions are normally granted contingent upon like permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$10 per page, table, or figure of APA material. Permission and fees are waived for authors who wish to reproduce their own material for personal use; fees only are waived for authors who wish to use their own material commercially. Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. copyright law: (1) those post-1977 articles that carry a code at the bottom of the first page, provided the per copy fee indicated in the code is paid through the Copyright Clearance Center, P.O. Box 765, Schenectady, NY 12301; (2) pre-1978 articles without fee. Instructors and educational institutions are permitted to photocopy isolated articles for noncommercial classroom or library reserve use without fee. Address requests for permission to the Permissions Office, APA, 1200 Seventeenth Street, N.W., Washington, D.C. 20036.

Subscriptions: Subscriptions are available on a calendar year basis only (January through December). Nonmember rates for 1978: \$32 domestic, \$34 foreign, \$6 single issue. APA member rate: \$16. Write to Subscription Section, APA.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

Back Issues and Back Volumes: For information regarding back issues or back volumes write to Order Department, APA.

Microform Editions: For information regarding microform editions write to any of the following: Johnson Associates, Inc., P.O. Box 1017, Greenwich, Connecticut 06830; University Microfilms, Ann Arbor, Michigan 48106; or Princeton Microfilms, Princeton, New Jersey 08540.

Change of Address: Send change of address notice and a recent mailing label to the attention of the Subscription Section, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee second-class forwarding postage.

Published bimonthly (beginning in February) in one volume per year by the American Psychological Association, Inc., 1200 Seventeenth Street, N.W. Washington, D.C. 20036. Printed in the U.S.A. Second-class postage paid at Washington, D.C., and at additional mailing offices.

Date..... 24-12-79.....
No. Librarian.....
Bureau Ednl Pay. Research.

Journal of Consulting and Clinical Psychology

Copyright © 1978 by the American Psychological Association, Inc.

October 1978 Volume 46, Number 5

Articles

- 839 Personality Traits and Environmental Variables as Independent Predictors of Post-hospitalization Outcome/*George A. Clum*
- 844 Rorschach Developmental Level and Intelligence Factors/*Roger P. Greenberg and Gilbert F. Cardwell*
- 849 Life Event Scales: Psychophysical Training and Rating Dimension Effects on Event-Weighting Coefficients/*Arthur A. Stone and John M. Neale*
- 854 Mood, Pleasant Events, and Unpleasant Events: Two Pilot Studies/*Lynn P. Rehm*
- 860 Cognitive and Personality Factors in Suicidal Behavior/*Andrew Geller and Alvin Atkins*
- 869 Relationships Between WISC-R Factors, Wide-Range Achievement Test Scores, and Visual-Motor Maturation in Children Referred for Psychological Evaluation/*James M. Stedman, G. Frank Lawlis, Robert H. Cortner, and Gloria Achterberg*
- 873 Freedom of Choice and Behavioral Change/*Frederick H. Kanfer and Laurence G. Grimm*
- 879 Reduction of Autonomic Arousal in Alcoholics: A Comparison of Relaxation and Meditation Techniques/*Jerry C. Parker, Gary S. Gilbert, and Richard W. Thoreson*
- 887 Sex and Worker Acceptance of a Former Mental Patient/*Amerigo Farina, Pauline J. Murray, and Thomas Groh*
- 892 Prospects for Faking Believable Deficits on Neuropsychological Testing/*Robert K. Heaton, Harold H. Smith, Jr., Ralph A. W. Lehman, and Arthur T. Vogt*
- 901 Validity of Self-Reports in Three Populations of Alcoholics/*Linda C. Sobell and Mark B. Sobell*
- 908 Drug Preferences of Multiple Drug Abusers/*Robert J. Harford*
- 913 Neuropsychological Stability in Multiple Sclerosis/*Robert J. Ivnik*
- ✓924 The Generalized Expectancy for Success Scale—A New Measure/*Bobbi Fibel and W. Daniel Hale*
- 932 Assessing the Impact of Life Changes: Development of the Life Experiences Survey/*Irwin G. Sarason, James H. Johnson, and Judith M. Siegel*
- 947 Young Adult Schizophrenics: Prediction of Outcome and Antecedent Childhood Factors/*James D. Roff and Raymond Knight*
- 953 Blood Alcohol Level Discrimination by Alcoholics: The Role of Internal and External Cues/*David Lansky, Peter E. Nathan, and David M. Lawson*
- 961 Cuban Value Structure: Treatment Implications/*José Szapocznik, Mercedes Arca Scopetta, Maria de los Angeles Aranalde, and William Kurtines*
- 971 Personality Characteristics of Long-Term Recovered Alcoholics: A Comparative Analysis/*William M. Kurtines, Leah R. Ball, and Gloria H. Wood*
- 978 Cognitive Preparation and Coping Self-Talk: Anxiety Management During the Stress of Flying/*Michel Givodo and Julius Roehl*
- 990 Use of Electromyographic Biofeedback and Cue-Controlled Relaxation in the Treatment of Test Anxiety/*D. Kenneth Counts, James G. Hollandsworth, Jr., and John D. Alcorn*
- 997 Vicarious Anxiety: A Systematic Evaluation of a Vicarious Threat to Self-Esteem/*Philip C. Kendall, A. J. Finch, Jr., and L. E. Montgomery*
- 1009 Illusory Correlation: A Further Exploration of Chapman's Paradigm/*Richard M. Kurtz and Sol L. Garfield*
- 1016 Avoidance-Approach: The Fifth Basic Conflict/*Seymour Epstein*

(Continued on next page)

- 1023 Effective Ingredients in Psychotherapy: Prediction of Outcome From Process Variables/*Beverly Gomes-Schwartz*
- 1036 Coping and the Self-Control of Chronic Tension Headache/*Kenneth A. Holroyd and Frank Andrasik*
- 1046 Comparison of Linear and Configural MMPI Diagnostic Methods With an Uncontaminated Criterion/*Ronald A. Giannetti, James H. Johnson, Daniel E. Klinger, and Thomas A. Williams*
- 1053 Longitudinal Study of Marital Success and Failure/*Peter M. Bentler and Michael D. Newcomb*
- 1071 Sum of MMPI Scales F, 4, and 9 as a Measure of Aggression/*L. Rowell Huesmann, Monroe M. Lefkowitz, and Leonard D. Eron*
- 1079 A Longitudinal Study of Coping Styles in Self-Defining and Socially Defined Women/*Abigail J. Stewart*
- 1085 Self Versus Ideal Self: A Comparison of Five Adjective Check List Indices/*Harrison G. Gough, Renato Lazzari, and Mario Fioravanti*
- 1092 Skill Training With Alcoholics/*Edmund F. Chaney, Michael R. O'Leary, and G. Alan Marlatt*
- 1105 Evaluating Alcoholism Treatment Programs: An Integrated Approach/*Ruth C. Cronkite and Rudolf H. Moos*
- 1120 Therapist Experience, General Clinical Ability, and Treatment Outcome in Schizophrenia/*A. Hussain Tuma, Philip R. A. May, Coralee Yale, and Alan B. Forsythe*
- 1127 An MMPI Scale to Separate Brain-Damaged From Functional Psychiatric Patients in Neuropsychiatric Settings/*Charles G. Watson and Duane Plemel*
- 1133 "Biosyntonic" Therapy: Modification of an Operant Conditioning Approach to Pedophilia/*J. Dennis Nolan and Curt Sandman*

Brief Reports

- 1141 Female Criminal Violence and Differential MMPI Characteristics/*Patricia B. Suther, Albert N. Allain, and Scott Geyer*
- 1144 Turning on or Turning Off: Sensation Seeking or Tension Reduction as Motivational Determinants of Alcohol Use/*Raymond M. Schwarz, Barry R. Burkhart, and Samuel B. Green*
- 1146 Therapist and Client Perceptions of Alternative Roles for the Facilitative Conditions/*Robert B. Slaney*
- 1148 Positive Versus Negative Self-Monitoring in the Self-Control of Smoking/*David A. Kantorowitz, Joyce Walters, and Kathy Pezdek*
- 1151 Aspects of the Self-Concept Related to Level of Self-Esteem/*Kenneth W. Christian*
- 1153 Changes in Preferences for Male and Female Counselors/*Elaine F. Walker and Jayne E. Stake*
- 1155 Imaging Vividness and the Outcome of in Vivo and Imagined Scene Desensitization/*John M. Dyckman and Philip A. Cowan*
- 1157 Psychophysiology of Fear Imagery: Differences Between Focal Phobia and Social Performance Anxiety/*Theodore C. Weerts and Peter J. Lang*
- 1160 Correlations and Factor Analysis of the WISC-R and the Peabody Picture Vocabulary Test for an Adolescent Psychiatric Sample/*Allan DeHorn and Valerie Klinge*
- 1162 Suicide and Transparency Responses on the Rorschach: A Replication/*Jill Rierdan, Elizabeth Lang, and Sara Eddy*
- 1164 Premorbid Social Competence Construct Generalizability Across Ethnic Groups: Path Analyses With Two Premorbid Social Competence Components/*Raymond M. Costello*
- 1166 Effects of the Sex of Both Interviewer and Subject on Reported Manifest Dream Content/*Ross B. Kremsdorf, Lucy J. Palladino, Douglas D. Polen, and Barbara J. Antista*
- 1168 Androgyny and Self-Esteem in the Upper Middle Class: A Replication of Spence/*Kevin O'Connor, David W. Mann, and Judith M. Bardwick*
- 872 Instructions to Authors

Personality Traits and Environmental Variables as Independent Predictors of Posthospitalization Outcome

George A. Clum

Virginia Polytechnic Institute and State University

A set of predictor variables, identified as intrapsychic, and a second set, identified as environmental, were examined in a multiple regression analysis as to their independent contribution in predicting posthospitalization adjustment. The analyses indicated that level of adjustment at baseline hospitalization was the most salient prognostic variable. Significant others' expectations for the patients' self-help performance contributed independently to a follow-up criterion of total symptomatology as rated by the significant other. The results provide moderate support for the hypothesis that environmental as well as intrapsychic variables are important prognostic indicators.

A recent review of prognostic factors of hospitalized psychiatric patients (Clum, 1975b) has shown that two types of variables, intrapsychic and environmental, can be hypothesized to be independent predictors of posthospital outcome. This conclusion was largely inferential, however, and no substantive data exist that demonstrate the independent contribution of each set of variables. To accomplish this, three conditions must be met: (a) A set of predictor variables, identified as intrapsychic, must be found to *predict* posthospital outcome; (b) a set of predictor variables, identified as environmental, must be found to predict posthospital outcome; and (c) both sets must contribute independent variance to the criterion in a multiple regression format. In line with the previous definition of these variables, intrapsychic variables include any "measurable cognitive or personality characteristics the patient exhibits. Environmental variables are defined as socio-

psychological phenomena within the patient's life space, but external to the patient himself" (Clum, 1975b, p. 416).

One example that may clarify the problem involves the relationship between marital status and prognosis. The personality trait view suggests that individuals who are single and socially withdrawn will not be selected as marriage partners and will tend to have a bad prognosis. Being single has been found to be related to an individual's continued stay in a mental institution and also to poor social adjustment on release from the hospital (Clum, 1975b). In contrast, the environmental view suggests that being married acts as a buffer that leads to a greater likelihood of the patient having a good prognosis. People who are married have a definite role in the family, whereas people who are single are neither the homemaker nor the breadwinner and hence have roles secondary to the functioning of the family. The press to perform will accordingly be less for the single individual and might be anticipated to lead to poorer performance for such people according to an expectation hypothesis.

To determine whether the relationship between single status and poor prognosis is due to a selection process or to a buffer hypothesis, the initial level of patient dysfunction must be controlled. If it can be shown that single individuals do not exhibit more disturbance

This study was supported by Grant UVAC-12-70 from the Commonwealth of Virginia Department of Mental Hygiene and Hospitals; by funds from the Fluid Research Fund, University of Virginia School of Medicine; and by a University Computing Grant, University of Virginia.

Requests for reprints should be sent to George A. Clum, Department of Psychology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24060.

than married individuals on admission to a hospital but that they do exhibit a poorer prognosis, the buffer hypothesis would be supported.

Another example that can be addressed to the question of independent contribution to prognosis would be that of expectations and outcome. The personality trait view would predict that the more disturbance exhibited by the individual, the lower would be the expectations, resulting in a poor prognosis; that is, level of disturbance would determine level of expectations. The environmentalist would predict outcome from expectations independent of level of disturbance.

The relative importance of these two sets of variables has been studied previously in relation to in-hospital criteria of length of hospitalization and rated improvement at discharge (Clum, 1975a). Although Clum's (1975a) study supported the hypothesis that environmental variables are predictive of length of hospitalization, it must be considered an inconclusive test. Length of hospitalization is a relatively weak criterion because of its fluctuation on the basis of administrative considerations and because of the fact that in a short-stay hospital such as the one in his study, the criterion variance is severely truncated. Accordingly, the present study expands this analysis using a criterion of rated adjustment 1 year after initial hospitalization.

Method

Subjects

Subjects in the present study ($N = 79$) included all patients admitted to the University of Virginia psychiatric service who were also available for follow-up evaluation 1 year after admission. This sample represented a severely truncated subsample (50%) of the subjects who had completed all baseline data, a reduction of subjects which, however, is comparable to other studies using mailings and phone calls to obtain follow-up data. A comparison of the two groups, however, revealed no differences on variables of marital status, age, race, social class, and initial level of symptomatology on five symptom factors. The subjects were between the ages of 16 and 65 and included only those who did not have a diagnosis of organic brain damage.

Predictors

At the time of hospitalization, each patient was administered a biographical inventory, the Life Change Inventory (described previously by Clum, 1976), and a measure of expectations for improvement. Information culled from the biographical inventory included total time previously spent in mental hospitals, education, income, marital status, age, and number of friends. The patients' expectations were derived from the Katz Adjustment Scales (Katz & Lyerley, 1963).

A significant other also rated his/her expectations for the patient's performance once the patient returned home from the hospital. Two measures of expectations were obtained—one regarding the patient's performance on self-help tasks, the other regarding the patient's social adjustment.

Criterion Measures

Two criteria were used—total number of symptoms as noted by the patient and total number of symptoms as noted by the significant other. The Katz Adjustment Scales were used to evaluate symptomatology.

Procedure

At the time of hospitalization, patients were asked to complete all research forms. A significant other (spouse, parent, relative, or friend) who was familiar with the patient's adjustment in the month prior to hospitalization was also asked to complete the research instruments. One year later the patient and the same significant other were again contacted, and the same forms were readministered. The sample was then divided randomly into validation and hold-out samples to determine the stability of the results.

Results

Patient-Related Symptomatology

Both environmental and personality trait measures were entered into a stepwise multiple regression equation in order to assess the independent contribution of each variable to a criterion of total symptomatology as rated 1 year after hospitalization. As Table 1 indicates, only total symptomatology at baseline, time in mental hospital, and education contributed independently to patient-rated total symptomatology at the 1-year follow-up. Further, only total symptomatology at baseline was found to contribute independently to the criterion in the hold-out cross-validation group.

Table 1
Regression Analysis of Predictor Variables at Hospitalization to Patient-Rated Total Symptomatology at 1-Year Follow-Up

Predictor variable	Sample			
	Validation		Hold out	
	Simple R	Beta weight	Simple R	Beta weight
Baseline patient-rated total symptoms ^a	.50 ^{b**}	.46	.47 ^{b*}	.28
Time in mental hospital (complex)	.26 ^b	.15	.41 [*]	
Education (complex)	-.39 ^{b*}	-.02	-.04	
Income (complex)	-.30		-.03	
Significant others' expectations—self-help ^c	.36 [*]		.32	
Marital status ^d	-.04		.04	
Age (complex)	-.21		-.12	
Patients' expectation—baseline ^a	-.15		-.03	
Level of performance as rated by significant other—social adjustment ^a	.03		.12	
No. friends (complex)	-.16		-.18	
Total stress ^a			.26	
Significant others' expectations—social adjustment ^a			.15	
Marital status			.14	
Single = 2	.02		.47	
All others = 1 (complex)	.69			

^a Intrapsychic.

^b Significant independent contribution in multiple regression.

^c Environmental.

^d Married = 1; all others = 2 (complex).

* $p = .05$.

** $p = .01$.

Stress was not found to be related to the criterion. However, significant others' expectations for the patient's performance on self-help tasks correlated in the expected direction with outcome but did not contribute independently to its predictions.

Significant-Other-Rated Symptomatology

The regression analysis of significant-other-rated total symptomatology at the 1-year follow-up (see Table 2) demonstrated that baseline total symptomatology, significant others' expectations for self-help, number of friends, marital status, and patients' baseline expectations were the only variables of those examined that contributed independently to the criterion. Significant others' expectations of self-help were significant at the .01 level, supporting the importance of expectations as a prognostic variable, independent of personality traits. Marital status was also reliably re-

lated to significant others' ratings of total symptomatology, again, in a negative direction. This was contrary to predictions. Of these five predictors only total baseline symptomatology and significant others' expectations cross-validated as independent contributors to the criterion.

The vast majority of previous research has found single people to have poorer prognoses. Since most of these studies were conducted on predominantly schizophrenic populations, it was decided to reanalyze the relationship between marital status and outcome separately for schizophrenic and schizoid personality patients and all other diagnostic groups. Pearson correlation coefficients were computed between marital status and total symptomatology at follow-up for patients diagnosed as schizophrenic or schizoid personality. The results indicated that marital status and total symptomatology at follow-up were related ($r = .41$ between marital status and significant others'

Table 2

Regression Analysis of Predictor Variables at Hospitalization to Significant-Other-Rated Total Symptomatology at 1-Year Follow-Up

Predictor variable	Sample			
	Validation		Hold out	
	Simple R	Beta weight	Simple R	Beta weight
Baseline significant-other-rated total symptoms ^a	.55 ^{b**}	.27	.40 ^{b*}	.36
Significant others' expectations—self-help ^c	.54 ^{b**}	.33	.24 ^b	.25
Marital status (complex) ^d	-.41 ^{b*}	-.45	-.33	
No. friends (complex)	.21 ^b	.27		
Patients' expectations—baseline ^a	.03 ^b	-.15	-.11	
Significant others' ratings of level of performance—social adjustment ^a	.47 ^{**}			
Age (complex)	-.08		-.08	
Education (complex)	-.31		-.23	
Significant others' expectations—social adjustment ^a	.17		.27	
Time in mental hospital (complex)	.22		.33	
Total stress ^c	.14		-.13	
Income (complex)	.17		-.14	
Marital status (complex)				
Single = 2	-.23 ^{**}	-.28	-.13	
All others = 1	.86		.59	

^a Intrapsychic.

^b Significant independent contribution in multiple regression.

^c Environmental.

^d Married = 1; all others = 2.

* $p = .05$.

** $p = .01$.

ratings; $r = .43$ between marital status and patients' ratings) in a generally positive direction, although not reaching significance due to the small sample size (8). Thus, for this subgroup of patients, a single status led to a poorer prognosis, as was previously predicted.

Discussion

The present study sought to determine whether environmental factors are of independent prognostic significance when considered conjointly with personality trait variables. A multiple regression format in which both types of predictors were considered was used to make this determination.

The results provided partial support for the hypothesis that predictor variables representative of both classes of variables would contribute independent predictor variance. Specifically, initial level of symptomatology was consistently related to the criterion. In addition,

significant others' expectations for self-help were related to total symptomatology as rated by the significant other but not as rated by the patient. This provides moderate support for the import of an environmental factor of others' expectations as an independent prognostic variable. Heretofore, it has been unclear whether the relationship of expectations to outcome was attributable to the fact that others' expectations were lower for those individuals whose adjustment was poor. It is arguable whether poor prognosis in this case was a function of the expectations or the already poor adjustment. The fact that expectations for self-help were not independently related to patients' ratings of symptomatology suggests that for this criterion the impact of expectations is determined largely by its relationship to initial level of symptoms.

No other prognostic variable, either personality or environmental, cross-validated.

This is especially confusing with regard to marital status, since this variable has been most consistently related to prognosis (Clum, 1975b). The vast majority of these studies, however, included Veterans Administration or state hospital patients, most of whom were schizophrenic and chronic. In the present study a marital status of married was positively correlated with significant-other-rated symptomatology, as compared with a negative relationship in most other studies. Since most of the patients in the present study were neither schizophrenic nor chronic, it was decided to analyze the relationship between marital status and outcome separately for those patients diagnosed as schizophrenic or schizoid and those with all other diagnoses. The results confirmed the finding of previous studies: Married status was related to a better outcome for the schizophrenic and schizoid subgroup. It is possible, therefore, that diagnosis is a moderator of the marital status-outcome relationship, rendering marital status an even more powerful prognostic variable. Whether it contributes independently to outcome for schizophrenic patients still remains an unanswered question.

The environmental variable of life stress was not found to be of any reliable consequence. However, in another study (Clum, 1976), it was demonstrated that whereas life changes prior to hospitalization were unrelated to symptomatology at follow-up, they were related to the level of symptoms at the time of hospitalization. Similarly, life changes subsequent to hospitalization were predictive of level of symptomatology 1 year after hospitalization. This supports Rahe's (1972) notion that stress only in the preceding year is predictive of symptomatology.

Implications

The implications of this study are twofold. First, the question of prognosis that confronts the clinician is related to decisions regarding disposition. For example, if a patient can be predicted to have a poor prognosis, the insistence on further follow-up treatment could

be increased. Decisions to continue hospitalization for a longer period of time or until such time as those factors predictive of outcome have been modified could be affected.

The second implication concerns those areas targeted for change. Since previous studies have focused on intrapsychic variables and have related these to outcome, the focus of the clinician has remained at effecting changes in the patient's personality. With the addition, in the present study, of the importance of expectations and possible role models as independent predictors of outcome, a new direction for targets for change should be forthcoming. Specifically, attention should be paid to changing the patient's families' expectations for performance, such that higher demands on the patients performance of self-help tasks should be emphasized. The importance of assuming a significant role in the family as a breadwinner or homemaker appears specific to schizophrenic patients. For this group, accordingly, the importance of developing significant social roles would seem to be a target for therapeutic intervention. In contrast, being married was found to be a negative prognostic factor for nonschizophrenic patients. Accordingly, affecting a more positive relationship within the marital context should receive greater attention in this group.

References

- Clum, G. A. Intrapsychic and environmental variables as predictors of length of hospitalization and rated improvement at discharge. *Journal of Consulting and Clinical Psychology*, 1975, 43, 276. (a)
- Clum, G. A. Intrapsychic variables and the patient's environment as factors in prognosis. *Psychological Bulletin*, 1975, 82, 413-431. (b)
- Clum, G. A. Role of stress in the prognoses of mental illness. *Journal of Consulting and Clinical Psychology*, 1976, 44, 1, 54-60.
- Katz, M. M., & Lyerley, S. B. Methods for measuring adjustment and social behavior in the community. I. Rationale, description, discriminative validity and scale development. *Psychological Reports*, 1963, 13, 503-535.
- Rahe, R. H. Subjects' recent life changes and their near future illness reports. *Annals of Clinical Research*, 1972, 4, 250-265.

Received October 18, 1976 ■

Rorschach Developmental Level and Intelligence Factors

Roger P. Greenberg

Department of Psychiatry
State University of New York
Upstate Medical Center, Syracuse

Gilbert F. Cardwell

DePaul University

This study sought to provide a more complete exploration of the relationship between Rorschach developmental level scores and intelligence than has appeared to date in the literature. Factor scores for both Wechsler Adult Intelligence Scale and Rorschach measures were used in comparing test protocols obtained from 86 psychiatric inpatients. Results reaffirmed a relationship between intellectual level and developmental level. Further, the relationship was found to be of a general nature. It did not rely specifically on verbal, perceptual, or memory factors. Results also indicated that the relationship held for females as well as males.

On the basis of Werner's (1948) developmental theory, Friedman (1953) derived a Rorschach measure of developmental level (DL). Goldfried, Stricker, and Weiner (1971) have reviewed a large number of the studies making use of this index and concluded that Friedman's scoring of the Rorschach results is a good measure of the DL of functioning. Findings indicate impressively stable and theoretically consistent age effects, theoretically consistent differences among pathological groups, correlations with the degree of regression among schizophrenics, and successful prediction of superior interpersonal functioning among high-scoring groups.

The relationship between DL and intelligence has been one issue of discussion in the literature. Werner (1948) speculated that mental development proceeds in the direction of increasing differentiation and hierarchical integration. This view offers a developmental framework for Ainsworth and Klopfer's (1954) hypothesis that "vague, global perception reflects a relatively low level capacity and . . . the more refined and differentiated

the perception the higher the level of intelligence" (p. 353). Thus, viewing the DL scoring of the Rorschach as an indicator of general cognitive functioning suggests a relationship with measures of intellectual ability.

In general, the literature has strongly supported the idea of such a relationship. For example, Goldfried (1962) found significant correlations between DL and IQ scores in a sample of male neuropsychiatric patients. Similarly, Blatt and Allison (1963) found IQ scores to be significantly related to developmentally high Whole Response scores in samples of graduate students and neuropsychiatric patients (Allison & Blatt, 1964; Blatt & Allison, 1963). Friedman and Orgel (1964) found a DL-IQ association in paranoid schizophrenics. They did not find the relationship in groups of catatonics, hebephrenics, neurotic brain-damaged patients, or normals. However, as Kissel (1965) has noted and demonstrated, the failure to find a relationship probably resulted from Friedman and Orgel's use of percentage scores rather than from an inherent lack of relationship between DL and intelligence.

Kissel (1965) did find the predicted relationship in a sample drawn from a child guidance clinic, and there has been a report of an association between DL and mental age in children (O'Neill, O'Neill, & Quinlan, 1976). Gerstein, Brodzinsky, and Reiskind

The assistance of the computer center at the Upstate Medical Center is gratefully acknowledged.

Requests for reprints should be sent to Roger P. Greenberg, Department of Psychiatry, State University Hospital, Upstate Medical Center, 750 East Adams Street, Syracuse, New York 13210.

(1976) also found the general relationship in a group of white children. Interestingly, however, in looking at children of below-average intelligence, they discovered that black subjects produced higher DL scores than did their white counterparts. They raised the possibility that a developmental-structural analysis of the Rorschach may provide a more realistic assessment of the cognitive capacity of some black children than do standard IQ tests.

Thus, there is now consistent evidence of a link between a Rorschach measure of cognitive development and tests of intellectual ability. However, since measures of intelligence are usually a composite of different kinds of abilities, Goldfried et al. (1971) suggested that it would be of value to specify which aspects of intellectual functioning are related to DL. The goal of the present article is to state more exactly the nature of the relationship between DL and intelligence. This goal is pursued by examining the relationship between two independent factors underlying DL and three independent factors underlying performance on the Wechsler Adult Intelligence Scale (WAIS). These factors are derived from previous factor analytic studies of the WAIS and a project of our own that examined the factor structure among Rorschach cards scored for DL.

Specifically, studies of the WAIS have extracted three factors in addition to a dominant general factor. These three dimensions have been labeled *Verbal Comprehension*, *Perceptual Organization*, and *Memory* (or *Freedom from Distractibility*) (Berger, Bernstein, Klein, Cohen, & Lucas, 1964; Cohen, 1957a, 1957b). Scores for the factors can be obtained by adding the scaled scores of the WAIS subtests making up each factor.

A construct validity study of the DL measure revealed two independent factors among Rorschach cards scored for DL (Cardwell & Greenberg, Note 1). One factor was the composite DL score obtained by adding up the subject's DL scores on each of the 10 Rorschach cards. The second factor was obtained by taking the difference between a subject's DL scores on two contrasting sets of four cards (i.e., DL score on Cards 3, 4, 5, and 6 minus the DL score on Cards 7, 8, 9, and 10).

In sum, the present study concerns the relationship between three distinct sources of variation within IQ data and two distinct sources of variation within the DL measure. This permits an examination of the total pattern of correlations among the five measures and a look at the number and form of dimensions that may contribute to covariation. If one predicts a nonspecific general factor underlying both DL and IQ, a relationship across all IQ factors and DL scores would be expected. This is the implication of Ainsworth and Klopfer's (1954) view. Alternatively, a more specific relationship would be suggested if DL dimensions were related to only verbal, perceptual, or memory factors on the WAIS.

In addition to providing a more thorough analysis of the relationship between IQ and DL, the present study also permits an examination of sex differences in the relationships. Previous studies of the DL construct have largely not included female subjects (Goldfried et al., 1971).

Method

Data used in the study were obtained from the file of a medical center psychology service. All Rorschach tests in the file had been administered and scored for DL by faculty members with extensive Rorschach experience or by psychology interns under the direct supervision of these faculty members. The interscorer reliability of the DL system has been found to be quite high, ranging from 89.7 to 95.5 (Friedman, 1953; Goldfried et al., 1971).

In selecting protocols for study, all referrals from the hospital's psychiatric inpatient service over a 7-year period were reviewed, and 86 records (57 women and 29 men) met the following criteria for study: (a) an available summary sheet listing developmental scoring for each card, (b) the production of at least one response to each card, and (c) an available summary sheet of WAIS subtest scores. The mean age of the sample was 29.1, with a range of 16 to 68 years. Subjects carried psychiatric diagnoses from all major categories (psychoses, personality disorders, psychoneuroses). Patients with a diagnosis of organic brain syndrome were excluded from the sample.

Friedman applied Werner's "orthogenetic principle" in the development of his scoring system. Werner (1957) holds that with increasing maturity, mental functioning "proceeds from a state of relative globality and lack of differentiation to a state of increasing differentiation, articulation, and hierarchic integration" (p. 126). Friedman's system, which attempts to measure such changes in perceptual func-

tioning, uses only structural and organizational aspects of the percept. Rorschach responses are first scored for location (whole or detail) and are then placed into one of six categories. The categories represent increasing levels of differentiation and integration. They are labeled (in order of increasing maturity) *amorphous* (*a*), *minus* (*-*), *vague* (*v*), *mediocre* (*m*), *plus* (*+*), and *plus-plus* (*++*) responses. A detailed presentation of the scoring criteria is presented by Friedman (1952, 1953) and Goldfried et al. (1971).

For the present study, each response given to the Rorschach was assigned a value between 1 and 12 according to the corresponding developmental level of maturity. All responses to a single card were then summed, and this total was divided by the number of responses given to that card. The weightings used for each response are as follows: $W_a = 1$; $D_a = 2$; $W_- = 3$; $D_- = 4$; $W_v = 5$; $D_v = 6$; $D_m = 7$; $W_m = 8$; $D_+ = 9$; $W_+ = 10$; $D_{++} = 11$; $W_{++} = 12$.

The order of the major categories amorphous through plus-plus follows directly from Friedman's interpretation of Werner. The relative rankings of whole and detail categories is consistent with Werner's orthogenetic principle and parallels the developmental trends reported by Friedman, Hemmendinger, and Siegel (cited in Phillips & Smith, 1953). We might also add that our 12-point ranking system represents a more differentiated and congruent quantitative translation of Friedman's scoring system than has been used before. For example, Becker (1956) used only a 6-point system by sometimes ignoring the distinction between whole and detail responses and by arbitrarily grouping minus, vague, and amorphous responses in some categories.

Consistent with the factor analytic studies of the WAIS (Berger et al., 1964; Cohen, 1957a, 1957b), three factor scores were derived by summing the total scaled scores of the appropriate WAIS subtests: Verbal Comprehension (Information, Comprehension, Similarities, and Vocabulary), Perceptual Organization (Picture Completion, Block Design, and Object Assembly), and Memory (Arithmetic, Digit Span, and Digit Symbol).

Results and Discussion

As an initial step in the analysis, covariance matrices were estimated for the five variables (three WAIS factors and two DL factors) separately for each sex. Box's (1949) technique was used in comparing these matrices, and this test produced an F ratio less than one, indicating little difference between the pattern of correlations for male and female subjects. This finding is in line with previous research, indicating that the DL construct can be meaningfully and consistently applied to both sexes (Cardwell & Greenberg, Note 1).

Table 1
Correlations of DL Factors with WAIS Factors and Full Scale IQ

WAIS factor	DL factor	
	1	2
Verbal	.40**	.35**
Memory	.28**	.21*
Perceptual	.34**	.28**
Full Scale IQ	.40**	.35**

Note. WAIS = Wechsler Adult Intelligence Scale; DL = developmental level.

* $p < .05$.

** $p < .01$.

The data were then combined across sexes, and the five-variable Pearson product-moment correlation matrix was estimated. The correlations between the three WAIS factors and the two DL factors from this matrix are presented in Table 1 along with the correlations between Full Scale IQ and the two DL factors.

As can be seen the correlations are all positive, significant, and of similar magnitude. Bartlett's (1941, 1947) test of complete independence between two sets of data was applied to correlations between WAIS and DL factors, and the result was significant, $\chi^2(6) = 44.74$, $p < .001$. The correlations between Full Scale IQ and the DL factors were also significant ($p < .01$). These results reaffirm previous research findings of a relationship between DL and intellectual measures.

To help answer questions concerning the specificity of this relationship, canonical correlations were estimated (Morrison, 1967). This technique attempts to identify sources of common variance between two sets of measures by estimating weighted combinations of the variables that are most strongly correlated. After one such dimension is specified, its effects are removed from the matrix of correlations. Additional canonical variables, which are independent of the first dimension, can then be identified among the residual correlations. Each time a dimension is extracted from the matrix, Bartlett's test of complete independence is reapplied to the residual correlation matrix until a nonsignificant result indicates that all common variance has been accounted for.

Table 2
Loadings for Canonical Roots, Canonical Correlations, and Significance Tests Between WAIS and DL Factors

Factor and test	Loadings for canonical roots	
	First	Second
WAIS		
Verbal	.761	.052
Memory	.474	-.725
Perception	.443	.686
DL		
1 (total score)	.869	.158
2	.684	.170
r	.611	.265
Bartlett's test (χ^2)	44.74**	6.59*

Note. WAIS = Wechsler Adult Intelligence Scale;

DL = developmental level.

* $p < .025$ ($df = 2$).

** $p < .001$ ($df = 6$).

Applied to the correlations between WAIS and DL factors, this technique produced the results summarized in Table 2.

A maximum of two canonical correlations is possible in these data, and both were significant. The first and most important dimension related, with essentially equal loading, the sum of the three WAIS factors to the sum of the two DL factors and produced a correlation of .61. This result strongly supports the notion of a nonspecific general factor as the most important source of covariance between WAIS and DL variables. Among WAIS variables, such a dimension appears very similar to Spearman's (1927) construct of "general intelligence (g).". Stern (1956) has conceptualized Spearman's g as a property of the mind that determines its capacity for "collective coupling" of separate intellectual factors. In essence, his discussion, much like discussions of the DL construct, focuses on an individual's ability to effectively integrate differentiated (mental) systems. Thus, differentiation and integration seem to be key concepts for both a theory of g and a theory of DL. As already noted, these concepts are also consonant with Ainsworth and Klopfer's (1954) view of a relationship between perceptual differentiation and level of intelligence.

In addition, it is not surprising to find that a measure based on Rorschach responses is associated with verbal abilities as well as perceptual skills. Indeed, numerous factor analytic studies of the Rorschach have shown that verbal abilities play a significant role in the style and quality of Rorschach responses given (Murstein, 1965).

The second canonical variable found in the present study accounted for substantially less common variance and produced a correlation of only .265. However, Bartlett's test revealed that this correlation is significant. The loadings indicate that within the patient sample studied, there was a tendency for high DL factor scores to be associated with high WAIS perceptual factor scores and low memory factor scores.

Overall, then, the present findings reaffirm a relationship between intellectual and psychological developmental factors in a psychiatric population. Other statements about the relationship can now also be made—namely, that it is largely of a general nature; it does not rely specifically on verbal, perceptual, or memory factors; and it applies to females as well as males.

Reference Note

1. Cardwell, G. F., & Greenberg, R. P. A multivariate analysis of the Rorschach developmental level score. Manuscript submitted for publication, 1977.

References

- Ainsworth, M. D., & Klopfer, B. Evaluation of intellectual level, control, creative potential, and the introversion-extraversion relationship. In B. Klopfer et al. (Eds.), *Developments in the Rorschach techniques* (Vol. 1). Yonkers-on-Hudson, N.Y.: World Book, 1954.
- Allison, J., & Blatt, S. J. The relationship of Rorschach whole responses to intelligence. *Journal of Projective Techniques and Personality Assessment*, 1964, 28, 255-260.
- Bartlett, M. S. The statistical significance of canonical correlations. *Biometrika*, 1941, 32, 29-38.
- Bartlett, M. S. Multivariate analysis. *Supplement to the Journal of the Royal Statistical Society*, 1947, 9, 176-197.
- Becker, W. C. A genetic approach to the interpretation and evaluation of the process-reactive dis-

- tion in schizophrenia. *Journal of Abnormal and Social Psychology*, 1956, 53, 229-236.
- Berger, L., Bernstein, A., Klein, E., Cohen, J., & Lucas, G. Effects of aging and pathology on the factorial structure of intelligence. *Journal of Consulting Psychology*, 1964, 28, 199-207.
- Blatt, S. J., & Allison, J. Methodological considerations in Rorschach research: The W response as an expression of abstractive and integrative strivings. *Journal of Projective Techniques*, 1963, 27, 267-278.
- Box, G. E. P. A general distribution theory for a class of likelihood criteria. *Biometrika*, 1949, 36, 317-346.
- Cohen, J. A factor-analytically based rationale for the WAIS. *Journal of Consulting Psychology*, 1957, 21, 451-457. (a)
- Cohen, J. The factorial structure of the WAIS between early childhood and old age. *Journal of Consulting Psychology*, 1957, 21, 283-290. (b)
- Friedman, H. Perceptual regression in schizophrenia: An hypothesis suggested by the use of the Rorschach test. *Journal of Genetic Psychology*, 1952, 81, 63-98.
- Friedman, H. Perceptual regression in schizophrenia: An hypothesis suggested by the use of the Rorschach test. *Journal of Projective Techniques*, 1953, 17, 171-185.
- Friedman, H., & Orgel, S. A. Rorschach developmental scores and intelligence level. *Journal of Projective Techniques and Personality Assessment*, 1964, 28, 425-428.
- Gerstein, A. I., Brodzinsky, D. M., & Reiskind, N. Perceptual integration on the Rorschach as an indicator of cognitive capacity: A developmental study of racial differences in a clinic population. *Journal of Consulting and Clinical Psychology*, 1976, 44, 760-765.
- Goldfried, M. R. Some normative data on Rorschach developmental level "card pull" in a psychiatric population. *Journal of Projective Techniques*, 1962, 26, 283-287.
- Goldfried, M. R., Stricker, G., & Weiner, I. B. *Rorschach handbook of clinical and research applications*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.
- Kissel, S. A brief note on the relationship between Rorschach developmental level and intelligence. *Journal of Projective Techniques and Personality Assessment*, 1965, 29, 454-455.
- Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill, 1967.
- Murstein, B. I. Factor analyses of the Rorschach test. In B. I. Murstein (Ed.), *Handbook of projective techniques*. New York: Basic Books, 1965.
- O'Neill, P., O'Neill, P. C., & Quinlan, D. M. Perceptual development on the Rorschach. *Journal of Personality Assessment*, 1976, 40, 115-121.
- Phillips, L., & Smith, J. G. *Rorschach interpretation: Advanced technique*. New York: Grune & Stratton, 1953.
- Spearman, C. *The abilities of man*. New York: Macmillan, 1927.
- Stern, A. W. The nature of g and the concept of intelligence. *Acta Psychologica*, 1956, 12, 282-289.
- Werner, H. *Comparative psychology of mental development* (Rev. ed.). New York: Follett, 1948.
- Werner, H. The concept of development from a comparative and organismic point of view. In D. B. Harris (Ed.), *The concept of development*. Minneapolis: University of Minnesota Press, 1957.

Received December 14, 1976 ■

Life Event Scales: Psychophysical Training and Rating Dimension Effects on Event-Weighting Coefficients

Arthur A. Stone and John M. Neale
State University of New York at Stony Brook

Subjects chosen randomly from the community ($N = 54$) responded to questionnaires and rated either "social readjustment" or "stressfulness" for the events in the Social Readjustment Rating Scale. Some subjects were given psychophysical training before rating the events. The training did not have a significant effect on the ratings. However, stressfulness ratings were consistently higher than those of social readjustment, and the data suggest that this effect interacted with the events rated. The interaction may mean that choice of event dimension influences the ability to predict a dependent measure from weighted life events.

The systematic scaling of life events, introduced by Holmes and Rahe (1967), was a major breakthrough in quantifying the relationship between life events and both physical and psychiatric illness. These investigators used Stevens' (1974) technique of magnitude estimation to determine the amount of "social readjustment" that certain events would cause for an "average" person, regardless of the desirability of the event. Each subject was given a three-paragraph explanation of the rating task and then estimated the magnitude of social readjustment that each event would require. The original event coefficients, so-called since they were used to multiplicatively weight the occurrence of events, were produced by a "sample of convenience" of 394.

Results of investigations using events weighted by Holmes and Rahe's (1967) coefficients are plentiful. Weighted event scores have been used to predict sudden cardiac arrest (Rahe & Lind, 1970), time of myocardial infarction (Rahe & Paasikivi, 1971; Theorell & Rahe, 1971), occurrence of bone fractures (Tollefson, 1972), scholastic achievement

(Harris, 1973), disease and illness rates (Gunderson & Rahe, 1974), slight colds and fevers (Holmes & Holmes, 1970), and depression (Paykel, 1973). Though only partial, this list illustrates the diverse usage that this method has recently enjoyed. The majority of these studies have demonstrated modest, positive correlations between life event scores and criterion measures.

In this article we address several issues concerning Holmes and Rahe's (1967) method. First, will a sample of subjects randomly selected from the community produce the same coefficients as the original sample? Second, given the brief instructions used by Holmes and Rahe, does practice with matching numbers to lines and lines to numbers as opposed to no training affect the coefficients? Third, does asking a subject to rate "stressfulness," a term that incorporates the concept of desirability rather than social readjustment, affect the ratings? If any of these variables are found to significantly affect the data, then we must address the question of what the implications are for investigators who must choose among the available sets of event weightings.

Method

Subjects

One hundred households were randomly selected from the county telephone directory. Each was then randomly assigned to one of four experimental conditions such that there were 25 households in each.

This research was supported in part by National Institute of Mental Health Grant MH 21145 and Office of Naval Research Grant N00D14-77-C-0693.

The authors thank Joan E. Broderick and James M. Waters for criticism of an early draft.

Requests for reprints should be sent to Arthur A. Stone, Department of Psychology, State University of New York, Stony Brook, New York 11794.

Table 1
Summary of Analysis of Variance on Logs of Raw Scores

Source	df ^a	F	<i>p</i>	
			F	Conservative F
Between subjects	53			
Rating dimension (A)	1	12.150	.002	.002
Training (B)	1	1.781	.189	.189
A × B	1	.204	>.300	>.300
Within subjects	2,173			
Events (C)	41	24.185	<.001	<.001
A × C	41	1.582	.012	.215
B × C	41	1.102	>.300	>.300
A × B × C	41	1.096	>.300	>.300
C × Subjects	2,050			
Total	2,267			

^a Not corrected for variance-covariance violations.

Households were sent a letter of introduction, a payment of \$2, and, depending on experimental assignment, one of four questionnaires. The enclosed directions stated that any adult member of the family was eligible to complete the form. Several follow-up reminders were sent if the questionnaire was not returned within 2 weeks. The final sample consisted of 54 people; all respondents save 1 were white, 76% were male, and most were considered middle class based on income and education.

All subjects completed magnitude estimation of the 43 items found on the Holmes and Rahe Social Readjustment Rating Scale (SRRS; Holmes & Rahe, 1967). The ratings were done in one of four conditions and formed a 2 × 2 factorial design. The first factor was whether the questionnaire incorporated a magnitude-estimation training procedure prior to filling out the SRRS. The training condition consisted of matching numbers to line lengths ranging from .3 cm to 30.0 cm and drawing lines to represent numbers ranging from 3 to 300. Each of these tasks was composed of two parts: a page with detailed instructions explaining how to use numbers and line lengths to represent relative magnitudes, followed by 13 number trials and 13 line trials. In the no-training condition, subjects completed the SRRS with no practice. Although the training procedure required that participants complete a longer form, return rates were not affected by this variable: 27 subjects in each group returned correctly completed questionnaires.

The second factor was the dimension that subjects used in rating the SRRS events. In one condition, social readjustment, the instructions were identical to those used by Holmes and Rahe (1967), that is, a three-paragraph description of the meaning of social readjustment, how to do magnitude estimations, and three examples of proportionally matching numbers to events. However, in the second con-

dition, stressfulness, the words *social readjustment* were replaced with *stressfulness*, and the word *change* was replaced with *stress*. One of the examples in the original SRRS instruction set was omitted from the instructions for the stressfulness condition, as there was no satisfactory means for changing the example so that it concerned stress and not social readjustment. One sheet of each questionnaire requested demographic information such as age, sex, education, income, and occupation.

Results

The hypotheses that training and rating dimension affect event coefficients were tested by a 2 (training vs. no training) × 2 (rating dimension: readjustment vs. stress) × 42 (the SRRS events) unweighted means, repeated measures analysis of variance (ANOVA). Data based on magnitude estimation procedures have repeatedly been demonstrated to conform to a log normal distribution (Stevens, 1974); thus, to meet the normality assumptions of the ANOVA, base-10 logarithms of raw scores were submitted to the analysis. Since departures from homogeneity of the variance-covariance matrices in repeated-measures ANOVAs produce positively biased *F* ratios (Box, 1954), the obtained *F* ratios were tested assuming both the worst case of the variance-covariance violations and assuming the case with no violations at all. If *F* ratios are significant with degrees of freedom adjusted for the

Table 2

Geometric Means on Social Readjustment Rating Scale Events for Stressfulness and Social Readjustment Conditions Collapsed over Training Factor

Item	Geometric <i>M</i> rating		Item	Geometric <i>M</i> rating	
	Stressfulness	Social readjustment		Stressfulness	Social readjustment
Trouble with boss	222	184	Change in church activities	160	95
Jail term	989	540	Marital reconciliation	716	375
Death of spouse	1,905	916	Fired from work	755	286
Change in sleeping habits	150	164	Divorce	1,151	438
Death of close family member	1,194	438	Change in line of work	420	292
Change in eating habits	135	150	Change in number of arguments	569	399
Foreclosure on mortgage	649	310	Change in work responsibilities	406	215
Change in personal habits	164	156	Wife beginning work	310	207
Death of close friend	804	252	Change in working conditions	366	181
Minor violations of law	111	66	Change in recreation	247	173
Personal achievement	261	227	Mortgage over \$10,000	389	231
Pregnancy	367	394	Mortgage less than \$10,000	230	121
Change in health of family member	687	342	Personal illness	675	421
Sexual difficulties	611	338	Business readjustment	581	247
In-law troubles	265	200	Change in social activities	226	154
Change in family get-togethers	161	112	Change in living conditions	361	237
Change in financial state	617	262	Retirement	347	332
Gaining a new family member	433	259	Vacation	189	117
Change in residence	296	158	Christmas	192	90
Child leaving home	403	226	Change to new school	215	141
Marital separation	1,219	443	Begin schooling	237	161

worst violation, then we have confidence in the effect. If an *F* ratio is not significant with this conservative test, yet is significant when tested under the no-violation condition, we are less confident in the effect and need more information to make any conclusive statements (Greenhouse & Geisser, 1959).

Table 1 presents the summary of the ANOVA with both uncorrected and conservative probabilities for each *F* ratio. The hypothesis that psychophysical training affects ratings of events was not confirmed, $F(1, 50) = 1.78$, *ns*. The significant main effect of rating dimension, $F(1, 50) = 12.15$, $p = .002$, confirms the hypothesis that ratings of

stressfulness were different from ratings of social readjustment. For all but three of the events rated, coefficients produced in the stressfulness condition were equal to or higher than the ratings of social readjustment. A significant main effect for events with a very high *F* ratio, $F(41, 2050) = 24.18$, $p < .001$, was expected. This simply indicates that different events were given different ratings. A significant Rating Dimension \times Events interaction was also observed, $F(41, 2050) = 1.58$, $p = .012$; but because the interaction is not significant under the conservative test, $F(1, 41) = 1.58$, $p = .215$, we are less confident in the effect. Geometric

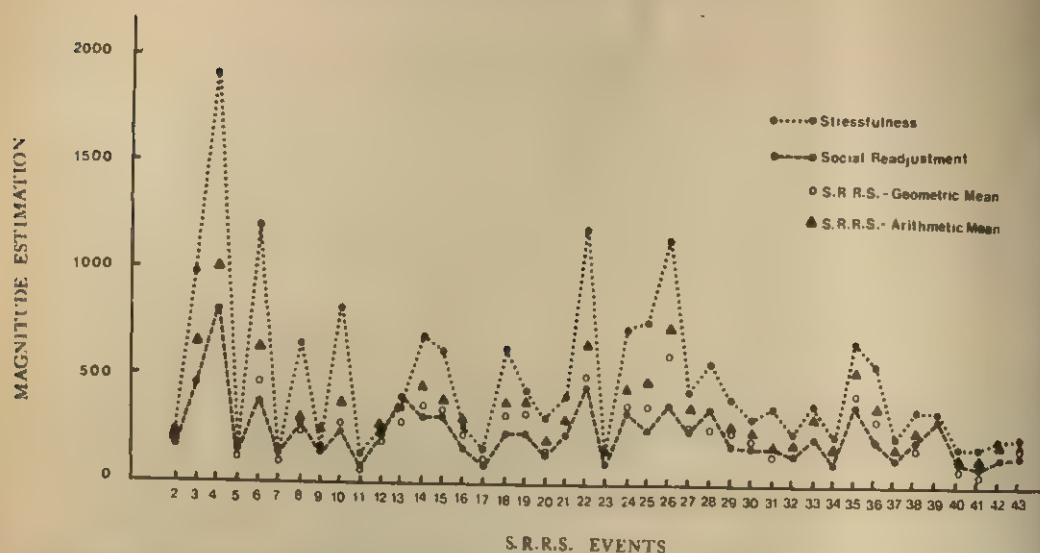


Figure 1. Relationship between arithmetic and geometric means of Social Readjustment Rating Scale (SRRS) weightings and geometric means of stressfulness and social readjustment conditions.

means for each of the rating dimensions, averaged across the training factor, are presented in Table 2.

Event coefficients obtained in the social readjustment condition, collapsed over levels of the training factor, are similar in magnitude to the coefficients that Holmes and Rahe (1967) found for the SRRS. Figure 1 presents the coefficients of the social readjustment condition found by us and Holmes and Rahe, as well as those obtained in the stressfulness condition of the present study. The observed differences between Holmes and Rahe and the present study's coefficients obtained in the social readjustment condition are largely a function of the measure of central tendency used to represent the data (see Figure 1). Thus, the social readjustment condition of the present study replicates the data that Holmes and Rahe obtained.

Discussion

To our surprise, trained subjects did not produce coefficients significantly different from those produced by the nontrained subjects. Thus, it is plausible that subjects can do magnitude estimations with little training, though Stevens (1974) believes that some practice may be useful. At least one

other interpretation of this finding must, however, be noted—The magnitude-estimation training instructions may not have been effective. Some support for this hypothesis lies in our observation that several subjects apparently used a straightedge while completing the training. Since the instructions did not explicitly prohibit this, an oversight on our part, it is possible that this occurred frequently. If this was the case, we must conclude that the training hypothesis was not adequately tested.

Coefficients from the stressfulness condition were shown to deviate significantly from the coefficients in the social readjustment condition in both elevation and configuration over events. But what do these differences mean to the life events researcher? Scale differences must be evaluated in light of the original purpose of weighting the events with these ratings; namely, is illness contraction related to the experience of life events? In predicting various illnesses from measures of life events, a single life events score is usually computed as a weighted linear combination of the events experienced. The weights are determined by subjects' ratings of events along some semantic dimension. Therefore, different rating dimensions would be expected to alter the weights used in forming

the linear combination. Changing these weights is likely, in turn, to alter the correlation between the life events score and the criterion, particularly if the relationship between the weights from the two rating dimensions is *not* monotonic and linear. Given the obtained interaction between rating dimension and events, we expect that correlations between events and an illness criterion will vary depending on the dimension used.

Our data demonstrate that the set of weights that an investigator uses may be linked to the prediction of criterion scores. Unfortunately, the direction of this effect is unknown. The researcher is thus left in the unenviable position of choosing among scales. An obvious and somewhat tempting resolution of this dilemma is for the investigator to use *all* sets of coefficients and empirically, for example, using a multiple regression analysis, determine the best combinations of scales for the set of data. One must wonder, though, what is to be learned about the perception of life events and their relationship to the criterion from such an exercise. Stated differently, the meaning of the weights yielded by a multiple regression is not tied to any substantive dimension of the events, such as social readjustment or stressfulness. Because the meaning of the prediction is of basic importance, identification of the factors that subjects use to rate life events is necessary for the advancement of an understanding of the interaction of life events with illness and psychological functioning.

We suggest systematic investigation of the properties or dimensions that subjects incorporate in their ratings. Multidimensional scaling methods are available for both testing the validity of rationally constructed dimensions and for deriving dimensions with no preconceptions. Nelson (1967) has similarly suggested that multidimensional analyses, in particular discriminant and factor analysis, be used for evaluating life events. Explorations along these lines have the potential to greatly clarify the complex inter-

actions that exist between person and environment.

References

- Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. *Annals of Mathematical Statistics*, 1954, 25, 290-302.
- Greenhouse, S. W., & Geisser, S. On methods in the analysis profile data. *Psychometrika*, 1959, 24, 95-112.
- Gunderson, E. K. E., & Rahe, R. H. (Eds.). *Life stress and illness*. Springfield, Ill.: Charles C Thomas, 1974.
- Harris, P. W. *The relationship of life change to academic performance among selected college freshmen at varying levels of college readiness* (Doctoral dissertation, East Texas State University, 1972). *Dissertation Abstracts International*, 1973, 33, 6665A-6666A. (University Microfilms No. 73-14, 285)
- Holmes, T. S., & Holmes, T. H. Short-term intrusions into life-style routine. *Journal of Psychosomatic Research*, 1970, 14, 121-132.
- Holmes, T. H., & Rahe, R. H. The Social Readjustment Rating Scale. *Journal of Psychosomatic Research*, 1967, 11, 213-218.
- Nelson, P. D. Comment. In E. K. E. Gunderson & R. H. Rahe (Eds.), *Life stress and illness*. Springfield, Ill.: Charles C Thomas, 1967.
- Paykel, E. G. Life events and acute depression. In J. P. Scott & E. C. Senay (Eds.), *Separation and depression: Clinical and research aspects*. Washington, D.C.: American Association for the Advancement of Science, 1973.
- Rahe, R. H., & Lind, E. Psychosocial factors and sudden cardiac death: A pilot study. *Journal of Psychosomatic Research*, 1970, 15, 19-24.
- Rahe, R., & Paasikivi, J. Psychosocial factors and myocardial infarction, II. An outpatient study in Sweden. *Journal of Psychosomatic Research*, 1971, 15, 33-39.
- Stevens, S. S. *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley, 1974.
- Theorell, T., & Rahe, R. H. Psychosocial factors and myocardial infarction, I. An inpatient study in Sweden. *Journal of Psychosomatic Research*, 1971, 15, 23-31.
- Tollefson, D. J. *The relationship between the occurrence of fractures and life crisis events*. Unpublished master's of nursing thesis, University of Washington, 1972.

Received February 9, 1977 ■

Mood, Pleasant Events, and Unpleasant Events: Two Pilot Studies

Lynn P. Rehm
University of Pittsburgh

The relationship between mood and both pleasant and unpleasant events is assessed in two studies. Undergraduates made self-ratings of mood and kept daily logs of pleasant and unpleasant events for approximately 2 weeks. Intrasubject correlations in both studies suggested that mood was related to pleasant and unpleasant events independently. Intersubject correlations were consistent but nonsignificant. Cross-lagged correlations were significantly less than same-day correlations. Weighted event scores produced marginally higher correlations with mood than unweighted scores. Minor sex differences are noted. The implications of these results for theory and practice are discussed.

A common assumption in the behavioral literature on depression is that mood is a function of reinforcement (e.g., Ferster, 1973; Lewinsohn, 1974). Although some basic questions relating to the precise definition of reinforcement in daily life remain unanswered, empirical relationships have been obtained between mood and pleasurable events or activities that have a likely correspondence to reinforcement. Using the Pleasant Events Schedule (MacPhillamy & Lewinsohn, Note 1), a 320-item list empirically developed to assess potentially reinforcing events in daily experience, intraindividual correlations have been demonstrated between mood and number of pleasant events (Lewinsohn & Graf, 1973; Lewinsohn & Libet, 1972). Depressed persons have been found to report fewer pleasant events than nondepressed psychiatric and normal groups (MacPhillamy & Lewinsohn, 1974). Wener and Rehm (1975) demonstrated a causal relationship between manipulated rate of positive feedback in a laboratory task and subsequent mood.

Self-monitoring of mood and pleasant events has been a part of a number of treatment programs for depression with a variety of rationales and functions. Lewinsohn's (1974) clinical research program employed the Pleasant Events Schedule as a means of empirically selecting targets for behavioral intervention. Events that correlate with mood for a specific individual are increased in order to influence mood (Lewinsohn, 1976). Fuchs and Rehm (1977) developed the Positive Activities Schedule, consisting of 20 categories of instrumental behavior likely to be associated with reinforcement. Depressed subjects kept a log of their daily activities using the schedule as a guide. Logs were used as a basis for self-selection of target behaviors and were also assumed to be an intervention in and of themselves, that is, an intervention modifying depressive, pessimistic self-monitoring.

Anton, Dunbar, and Friedman (1976) used activity logs as part of a therapy program that included scheduling of individual reinforcing activities. Enjoyability ratings of each activity logged were described as a potential dependent variable. Ad hoc activity schedules were used in three case studies described by Rush, Khatami, and Beck (1975). Activity data were used to confront clients' cognitively distorted interpretation of their behavior. Another ad hoc use of activity logs

The writing of this article was supported in part by National Institute of Mental Health Grant MH27822.

Requests for reprints should be sent to Lynn P. Rehm, Clinical Psychology Center, 606 Old Engineering Hall, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

Table 1
Average Mood and Events Data Across Subjects: Study 1

Variable	Males ^a		Females ^b		Total	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Mean daily mood rating	4.91	.76	5.99	.76	5.63	.91
Mean daily pleasant events	3.66	1.05	4.12	1.66	3.97	1.48
Mean daily unpleasant events	2.85	1.36	2.51	1.28	2.62	1.29

^a *n* = 10.

^b *n* = 20.

was described by McLean (1976), who used them primarily to assess improvement in behavioral productivity as one alternative component of therapy for depression.

Given the widespread use of pleasant event or activity monitoring in depression therapy programs and the empirical support for the relationship between pleasant events and mood, it is somewhat surprising that only activities associated with reward have been studied. All of the research cited has dealt with "pleasant events" or "positive activities," yet aversive events have been related to depression in a number of ways. Aversive, stressful life events have been found to precede clinical depression (cf. Dohrenwend & Dohrenwend, 1974). Seligman (1975) has demonstrated empirically that noncontingent aversive events lead to a state of learned helplessness, which he associates with depression. Lewinsohn, Lobitz, and Wilson (1973) found that depressed persons were particularly sensitive to aversive events. It would seem logical that mood would be related to aversive events in the daily life of normal persons.

Two pilot studies were conducted to explore the feasibility of assessing the relationship between unpleasant events and mood in a manner that could be adapted to clinical usage. In both studies events were listed by individuals in positive or negative columns on a daily log with a general definition as a guide. Pleasant events were defined as any event that is pleasant, enjoyable, or rewarding, and unpleasant events were defined as any event that is unpleasant, aversive, or punishing. Mood was assessed on a daily basis on a 0-10 rating scale, anchored at 0 = worst

mood ever to 10 = best mood ever. This scale was adopted for simplicity and ease of administration. Aitken (1969; Aitken & Zealley, 1970) described the use of a similar simple scale that correlated well with psychiatric ratings of depression. In both studies subjects were instructed to make their mood ratings and then fill out the events logs daily at the end of the day.

Study 1

Method

Log sheets and standard instructions were given as a class assignment to 33 undergraduates enrolled in a psychology course. Usable data were obtained from 30 subjects (10 males and 20 females) who kept logs for approximately 2 weeks (1 for 13 days, 13 for 14 days, 15 for 15 days, and 1 for 16 days).

Results

The actual events recorded varied somewhat in specificity and type. Some sample pleasant events that were recorded by students included "read Sunday paper," "good lunch," "sunny day," "letter from Adrienne," "bought albums," "good grade on bio test," "Joni Mitchell concert," "party," "saw good movie," "got high," and "complimented by Mrs. F." Some examples of unpleasant events were "bio test," "dentist appointment," "got parking ticket," "missed bus," "got blister playing squash," "Pitt lost game," "argument with roommate," "did the laundry," and "dull Geography class."

Means for data collected are shown in Table 1. Males' ratings of their mood were significantly lower than those for females, $t(28) = 3.66$, $p < .01$. Differences in number

Table 2
Average Intraindividual Correlations: Study 1

Correlated variables	M				
	Males ^a	Females ^b	Total ^c	With previous events ^c	With subsequent events ^c
Mood and pleasant events	.65	.55	.58	.11	.15
Mood and unpleasant events	-.36	-.48	-.45	-.11	.00
Pleasant and unpleasant events	-.15	-.28	-.22	—	—
Mood and events (<i>R</i>)	.75	.68	.70	—	—

^a *n* = 10.

^b *n* = 20.

^c *n* = 30.

of events were not significant though in directions consistent with the mood differences.

Mean correlations were calculated using Fisher's *z* transformation and were then converted back to Pearson product-moment correlations. No significant differences in correlations were found between sexes. Looking at the total group, the average correlations between pleasant and unpleasant events with mood were all statistically significant. The relatively low, negative, average correlation between the two classes of events suggests that pleasant and unpleasant events make relatively independent contributions to mood. The average multiple correlations for pleasant and unpleasant events correlated with mood (see Table 2) were larger than either single correlation and this further suggests mood is a function of both pleasant and unpleasant events.

Lewinsohn and Libet (1972) posed the question of whether mood would be affected by the previous day's activity or whether mood would affect the subsequent day's activity. This question can be answered by looking at correlations between mood and events of the previous day and between mood and events of the subsequent day. Table 2 shows that these mean cross-lagged correlations are quite low and comparable to those obtained by Lewinsohn and Libet. The same-day correlations were significantly larger than those for mood and events of the previous day: For mood and pleasant events, $t(29) = 9.30$, $p < .01$; for mood and unpleasant events, $t(29) = 5.40$, $p < .01$, or for those for mood

and events of the subsequent day, for mood and pleasant events, $t(29) = 6.55$, $p < .01$; for mood and unpleasant events, $t(29) = 5.95$, $p < .01$. Thus there was no indication of causality from events of one day to mood on the next or from the mood of one day to the events of the next. Lewinsohn and Libet's finding for pleasant events and mood was replicated and extended to unpleasant events as well.

Although the methodology is directed toward intrasubject questions, intersubject data can also be derived. For the 30 subjects, average mood correlated .31 with average pleasant events and $-.28$ with average unpleasant events. The multiple correlation between mood and both event scores was .57. These data are consistent with the intrasubject data, though only the latter correlation is statistically significant ($df = 28$, $p < .01$). A correlation of .46 ($df = 28$, $p < .05$) between average pleasant and average unpleasant events suggests individual differences in list length for this self-monitoring format.

Study 2

The second study had three purposes: (a) to replicate the first study with regard to monitoring pleasant events; (b) to determine whether a value weighting for each event would yield an obtained reinforcement measure that would enhance correlations with mood; and (c) to determine whether perceived contingency contributes to mood correlations.

Table 3

Average Mood and Events Data Across 14 Subjects: Study 2

Variable	Raw score		Weighted score		Contingent score	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Mean daily mood ratings	5.83	.80	—	—	—	—
Mean daily pleasant events	5.53	1.96	32.52	12.51	19.33	7.66
Mean daily unpleasant events	3.68	1.71	21.00	10.60	11.22	8.27

Method

Thirty-four undergraduates enrolled in a psychology course were assigned a choice of projects, one of which was to participate in this study. Fourteen students elected to participate and kept monitoring data for 14 days. Instructions were the same as in Study 1, except for the addition of weighting and perceived contingency instructions. After recording each event, subjects were instructed to give it a value on a 0-10 scale. For pleasant events, 10 was to signify "an extremely enjoyable or pleasant event," and for unpleasant events, 10 was to signify "an extremely aversive or unpleasant event." Zero signified neutrality on either scale.

To assess perceptions of contingency, participants were asked to indicate how many of the value points assigned to each event were directly attributable to themselves or their own behavior (i.e., their effort or lack of effort or their skills or lack of skill). This method allowed for a more continuous assessment of degree of perceived response contingency than would a dichotomous yes or no. Attribution research (cf. Weiner et al., 1971) suggests that such judgments are continuous.

Results

The average daily mood rating for these subjects was 5.83, which corresponds closely with Study 1. Subjects in this study (see Table 3) recorded more events, both pleasant, $t(42) = 2.652$, $p < .05$, and unpleasant $t(42)$

$= 2.050$, $p < .05$, than subjects in Study 1. These differences may be due to a different set acquired from the class discussion of the project or to self-selection differences. In either case they do not effect the hypotheses in question.

Mean correlations between mood and events (see Table 4) were slightly smaller but comparable to those obtained in Study 1. Again, pleasant and unpleasant events were uncorrelated with each other on the average, but each contributed to mood variance.

The use of weightings enhanced the average intraindividual correlations to a minor degree. A comparison of pairs of correlations with mood between unweighted and weighted pleasant events scores indicated that the weighted correlations were larger to a marginally significant degree, $t(13) = 1.809$, $p < .05$, one-tailed. The increase in the mean unpleasant events correlation with weighting was not significant, $t(13) = 1.240$. Thus the evidence suggests that weighting event scores may have only a slight value in identifying relationships with mood more accurately.

The average correlations of mood and self-attributed points were of the same magnitude as the correlations between mood and

Table 4

Average Intrasubject and Intersubject Correlations: Study 2

Correlated variables	Intrasubject <i>M</i>			Intersubject <i>M</i>		
	Raw scores	Weighted scores	Contingent scores	Raw scores	Weighted scores	Contingent scores
Mood and pleasant events	.51	.59	.52	.12	.18	.09
Mood and unpleasant events	-.35	-.41	-.27	-.13	-.12	-.13
Pleasant and unpleasant events	.08	-.01	.03	.71	.62	.50
Mood and events (<i>R</i>)	.70	.75	.70	.33	.34	.22

number of events. No particular advantage appears to accrue from specifying reinforcement value that is perceived as contingent on the person's own behavior.

Again it is notable that the average correlation between pleasant and unpleasant events was close to zero. Each class of events appears to contribute to mood independently. Average multiple correlations between mood and both sets of event scores (Table 4) bear out this assertion.

As can be seen in Table 4, interindividual correlations between mood and events were in the predicted directions but were quite small and not statistically significant. Correlations between pleasant and unpleasant events scores were significant for raw scores ($p < .01$) and weighted scores ($p < .05$). These suggest significant individual differences in reporting rates for pleasant and unpleasant events even though these classes are relatively independent within subjects. Multiple correlations for mood with the events scores combined were again consistent with intrasubject findings but were not statistically significant.

It is particularly interesting that perceived contingent event scores do not enhance correlations with mood. As an additional way of examining whether perceived contingency is related to mood, a ratio of self-attributed points to total weighted scores was calculated for each subject for both pleasant and unpleasant events. Averaged across all individuals, 59.4% of the value of pleasant events was self-attributed, whereas 49.5% of the value of unpleasant events was self-attributed. Correlations between these ratios and average mood would indicate whether subjects who tend to attribute a greater proportion of their pleasant or unpleasant events value to themselves would be more or less depressed than subjects who see these events as less contingent. The correlation between the ratio of perceived contingent pleasant events and mood was $-.14$. The correlation between the ratio of perceived contingent unpleasant events and mood was $.004$. Needless to say, neither was significant. Thus there was no evidence that tendency to perceive contingency between events and behavior influenced mood in this study.

Discussion

These two studies suggest that even with this simple methodology, it is feasible to assess the relationships between mood and events, both pleasant and unpleasant, in meaningful ways. Despite individual differences in recording rates, pleasant and unpleasant events were recorded with relative independence, and each correlated with mood. The magnitude of correlations between mood and pleasant events using these less structured methods was comparable to that obtained by Lewinsohn and Libet (1972) and Lewinsohn and Graf (1973). Although the correlations with unpleasant events were regularly of a lesser magnitude, they nevertheless make a significant additional contribution to mood.

Evidence from the second study suggests that a simple scaling of event magnitude contributes only marginally to greater precision in identifying relationships between mood and events. Differentiating the value by the degree to which events are self-attributed did not seem to contribute to correlations with mood for either pleasant or unpleasant events. Events are related to mood whether contingent on one's own behavior or entirely external in origin.

Thus there appears to be a potential clinical value to self-monitoring unpleasant as well as pleasant events in conjunction with behavior therapy for depression. Unpleasant events that correlate with mood may also be appropriate target activities for modification (Lewinsohn, 1976). The data generated by monitoring unpleasant events may also contribute to helping clients make more accurate discriminations about functional relationships between their behavior and their mood. As a result, they may be able to make more realistic evaluations and set more realistic goals via techniques analogous to those used by Fuchs and Rehm (1977) or Rush et al. (1975).

Certain qualifications must accompany these conclusions. First, these studies should indeed be considered as only promising pilot studies. Problems concerning the precise definition and validity of unpleasant events remain, as they remain for pleasant event re-

cording. Further basic research is certainly necessary. Second, these studies suggest potential for monitoring unpleasant events only in terms of intrasubject investigations. Inter-subject correlations were largely nonsignificant; individual differences in recording rates were evident; and sex differences were obtained. The development of structured methods for assessing unpleasant events, parallel to those available for assessing pleasant events, may allow for valid nomothetic studies.

Reference Note

1. MacPhillamy, D., & Lewinsohn, P. M. *The Pleasant Events Schedule*. Unpublished manuscript, University of Oregon, 1971.

References

- Aitken, R. C. B. Measures of feeling using analogue scales. *Proceedings of the Royal Society of Medicine*, 1969, 62, 989-993.
- Aitken, R. C. B., & Zealley, A. K. Measurement of mood. *British Journal of Hospital Medicine*, 1970, 4, 214-224.
- Anton, J. L., Dunbar, J., & Friedman, L. Anticipation training in the treatment of depression. In J. D. Krumboltz & C. E. Thoresen (Eds.), *Counseling methods*. New York: Holt, Rinehart & Winston, 1976.
- Dohrenwend, B. P., & Dohrenwend, B. S. (Eds.). *Stressful life events: Their nature and effects*. New York: Wiley, 1974.
- Ferster, C. B. A functional analysis of depression. *American Psychologist*, 1973, 28, 857-870.
- Fuchs, C. Z., & Rehm, L. P. A self-control behavior therapy program for depression. *Journal of Consulting and Clinical Psychology*, 1977, 45, 206-215.
- Lewinsohn, P. M. A behavioral approach to depression. In R. M. Friedman & M. M. Katz (Eds.), *The psychology of depression: Contemporary theory and research*. New York: Wiley, 1974.
- Lewinsohn, P. M. Activity schedules in treatment of depression. In J. D. Krumboltz & C. E. Thoresen (Eds.), *Counseling methods*. New York: Holt, Rinehart & Winston, 1976.
- Lewinsohn, P. M., & Graf, M. Pleasant activities and depression. *Journal of Consulting and Clinical Psychology*, 1973, 41, 261-268.
- Lewinsohn, P. M., & Libet, J. Pleasant events, activity schedules and depression. *Journal of Abnormal Psychology*, 1972, 79, 291-295.
- Lewinsohn, P. M., Lobitz, W. C., & Wilson, S. Sensitivity of depressed individuals to aversive stimuli. *Journal of Abnormal Psychology*, 1973, 81, 259-263.
- MacPhillamy, D. J., & Lewinsohn, P. M. Depression as a function of levels of desired and obtained pleasure. *Journal of Abnormal Psychology*, 1974, 83, 651-657.
- McLean, P. Therapeutic decision-making in the behavioral treatment of depression. In P. O. Davidson (Ed.), *The behavioral management of anxiety, depression and pain*. New York: Brunner/Mazel, 1976.
- Rush, A. J., Khatami, M., & Beck, A. T. Cognitive and behavior therapy in chronic depression. *Behavior Therapy*, 1975, 6, 398-404.
- Seligman, M. E. P. *Helplessness: On depression, development and death*. San Francisco: Freeman, 1975.
- Weiner, B. et al. *Perceiving the causes of success and failure*. Morristown, N.J.: General Learning Press, 1971.
- Wener, A. E., & Rehm, L. P. Depressive affect: A test of behavioral hypotheses. *Journal of Abnormal Psychology*, 1975, 84, 221-227.

Received March 2, 1977 ■

Cognitive and Personality Factors in Suicidal Behavior

Andrew M. Geller
Yeshiva University

Alvin Atkins
Montefiore Hospital and Medical Center
New York, New York

The role of aggression in suicidal behavior is studied. The personality functioning of 20 suicide attempters, 20 nonsuicidal psychiatric controls, and 20 suicide completers was assessed using the Rorschach. There were 11 women and 9 men in each group, and their ages ranged from 21 to 63. Feffer's role-taking task provided a test of the cognitive functioning of the first two groups. All three groups experienced the breakthrough of more aggressive than libidinal drive derivatives, but no significant differences between the groups were found. Only the suicide attempters' aggressive responses were more primitive than their libidinal responses. On the role-taking task, the suicidal group's cognitive functioning in the neutral situation was superior to their functioning in the aggressive one. The control group yielded no such difference. The suicidal group's performance in the aggressive situation was also significantly inferior to the control group. These results are interpreted as underscoring the role of cognition in symptom choice.

Conflicts over aggression play a central role in the psychoanalytic theory of suicide. Freud (1917/1957) struggled with the problem of how the ego, with its self-love and narcissistic libido, could allow its self-destruction. The answer, he theorized, lies in the ego's treating itself as an object, directing against that object the hostility and sadism that were its original reaction to objects in the outside world. Freud (1920/1950) later added to his theory the concept of a drive toward death, to be found in the controlling, coercing, and punishing components of the superego. In depression, the superego obtains a hold on consciousness and, as a "pure culture of the death instinct," drives the ego into death.

Other psychoanalytic thinkers have modified this theory. Zilboorg (1937) wrote that suicide is likely only when the individual has identified with a dead person, and that the process of identification has to take place

during childhood or adolescence, at a time when the incorporated person is already dead. Menninger (1938) agreed with Freud that suicide is the wish to kill another turned inward, but he also saw it as the ego's punishing itself for that crime and wishing to be killed.

Although the early psychoanalytic literature on suicide was based primarily on case studies, later writers have attempted broader, more experimental analyses of the role of aggression in suicidal behavior. Their results are inconclusive. Several researchers found that suicidal individuals are more hostile than nonsuicidals (Vinoda, 1966), that their dreams contain more themes of violence than control subjects (Raphling, 1970), and that they tend to resent those on whom they depend (Lester, 1969). Others found no significant differences in aggressive impulses or ideation between suicidal and nonsuicidal subjects (Eisenthal, 1967; Fisher, 1971).

Little research has focused on the role of cognitive functioning in suicidal behavior, although there is evidence to suggest that cognitive dysfunction can lead to symptomatic behavior. Feffer (1959) extended Piaget's (1950) concept of decentering activity in the

This article is based on a doctoral dissertation submitted by Andrew M. Geller to Yeshiva University under the supervision of Alvin Atkins.

Requests for reprints should be sent to Andrew M. Geller, who is now with Hay Associates, 1 Dag Hammarskjöld Plaza, New York, New York 10017.

physical, inanimate world to an analysis of the individual's cognitive structuring of social content. In his theory, roles and role reciprocals are the social polarities in a relationship of subject and interpersonal object that parallel the subject-object relationship of impersonal cognition. In order for subjects to construct the reality of the other accurately, to be able to modulate their behavior in anticipation of the behavior of the other, they must be able to coordinate simultaneously their role and the role of the other in the same situation. They must be aware of both perspectives. Prior to attainment of this ability to simultaneously decenter in social situations, the subjects are able only to react to the behavior of the other. At yet a more primitive level, subjects have great difficulty even being aware of perspectives different from their own, and their behavior is almost exclusively egocentric.

Lowenherz and Feffer (1969) found that adult subjects had greater difficulty coordinating role perspectives when the role was one that was defensively isolated, that is, when the subject had defined one of the role reciprocals as "least like me" or "personally unacceptable to me." This elaboration of Feffer's model of interpersonal cognitive development has been used to explain symptomatic behavior. Such behavior may be thought of as the result of isolation between dynamically relevant schemata. The subjects' relationship to reality is distorted, as they focus on only one dimension in the situation, unaware of the corrective influence of the opposite perspective. For example, Ward (1976) tested the hypothesis that the "other" in psychotic delusions, particularly if that other is a malevolent, threatening figure, is a representation of defensively isolated aspects of the self. She found that the subjects indeed functioned at a more primitive cognitive level when taking the roles with delusional themes than they did when working with neutral themes.

The present study examined both cognitive and defensive functioning in seriously suicidal individuals. Given the central rule that aggression plays in the psychoanalytic theories of suicidal behavior, it was predicted that within the suicidal group there would be more break-

through of primary process elements in response to aggressive stimuli or impulses than in response to libidinal stimuli and impulses, and the defenses against these breakthroughs would be less effective for aggressive than for libidinal material. It was expected that as a result of these weak defenses, the thinking of the suicidal subjects would contain a greater proportion of primary process elements than would the thinking of the control group in dealing with aggressive stimuli and impulses. Finally, it was predicted that subjects whose symptomatic behavior included making serious suicide attempts would have difficulty coordinating the roles represented by that symptom, namely victim and aggressor. This difficulty has been interpreted as a function of the subject's more primitive cognitive functioning when assuming the role of the aggressor, which is the reciprocal of the symptomatically exaggerated role orientation of the victim.

Method

Subjects

There were two groups of subjects, one experimental and one control. Each group was comprised of 20 subjects with a wide range of age, education, and socioeconomic status. All subjects were inpatients in a private psychiatric hospital in New York City.

The first experimental group was made up of 20 randomly selected patients who had made serious suicidal attempts. Only subjects whose suicide attempts were potentially lethal were included in this group. Patients who met that criterion but at the time scheduled for testing were overtly psychotic (actively hallucinating and delusional) were excluded from the study. Also excluded were patients who were diagnosed as suffering from more than minimal brain damage, more than minimal retardation, or a life-threatening physical illness. Although each subject's diagnosis, at both admission and discharge, medication regimen, family history, sex, marital status, education, and socioeconomic level were noted, these factors played no part in determining eligibility for the study.

The control group, also comprised of 20 subjects, was made up of randomly selected inpatients at the same hospital. Each subject had no history of suicidal ideation, suicidal impulses, or gestures. Both the history of the subject's present illness and past history as recorded in the chart, plus consultation with the subject's therapist, were used in making this determination. As with the experimental group, patients who were actively psychotic or suf-

fering from serious organic brain disease, mental retardation, or a life-threatening physical illness were excluded.

A third "group" was composed of data from patients who had actually committed suicide. Twenty Rorschach protocols were randomly collected from records of subjects who were tested while inpatients at the hospital and who within 1 year of that testing killed themselves. The use in this study of the available data from this group is an attempt to partially eliminate a conceptual difficulty found in most research on suicide. In general, such studies have examined subjects who attempted suicide and maintained that the findings were also applicable to those individuals who succeed in killing themselves. In this study, the data from those who committed suicide were used as a comparison with the information obtained from the suicide attempter group, providing some notion of the validity of that information.

The demographic characteristics of all three groups were similar. The ages ranged from below 20 to over 60, with a majority of subjects between 20 and 50 years of age. There were 11 women and 9 men in each group, and all three groups were virtually all Caucasian with the exception of 2 black subjects. Over half of the subjects in each group were single. In the nonsuicidal group, over half were Catholic, but religion was fairly equally represented in the other groups. All subjects in each group were high school or college graduates.

Procedure

All testing was done in individual sessions held in an office on the subject's inpatient unit. Inclusion in the study was voluntary, and each subject was assured of confidentiality and anonymity.

There were two parts to the stimulus situation. The Rorschach was used to test the defensive functioning of the subjects in the presence of aggressive and nonaggressive stimulation. Feffer's (1959) role-taking task (RTT) was used to measure their cognitive functioning in aggressive and neutral situations. The two tests were presented to the subjects in counterbalanced order.

Rorschach

The Rorschach inkblots were administered to each subject in the usual fashion. Subjects were first shown the 10 cards in sequence and were asked to tell what they looked like or reminded them of. On the first two cards, if only one response was given, another was requested. Thereafter, no further encouragement was given by the examiner. After the subjects had responded to all 10 cards, they were asked to go through each one again with the examiner to explain what it was about the blot that determined the response given.

Subjects' responses to the Rorschach were scored according to Holt's (Note 1) system for scoring primary process on the Rorschach. For Holt, the

distinction between primary and secondary process is a function of the degree to which drive dominates logic and reality in a response and the nature of the drive itself. Holt's system measures manifestations of primary process on the Rorschach and the means by which a person tries to control it and defend against the anxiety its emergence presumably entails. Scorable Rorschach responses are divided into two groups, those motivated by libidinal drives and those resulting from aggressive drives. These two categories are further divided into 11 subgroups, such as oral, anal, aggression-subject, and aggression-object. Each response is scored 1 or 2 depending on its level of what Holt called "primariness." The arbitrary boundaries between the two levels are defined by two points. One is a continuum from raw, shocking, blatant, or primitive forms of the drive in question (Level 1) to civilized, socially acceptable forms that are more appropriate to social communication between strangers in a professional situation (Level 2). The second criterion is the degree to which the response focuses on the drive-relevant aspect of an image, a division useful primarily for libidinal responses.

Because this binary division seemed too crude, Holt introduced a 5-point scale of Defense Demand (DD), which serves primarily to further differentiate responses along a continuum of primariness, with the DD score rising from 1 to 5 as the response becomes less socially acceptable and more bizarre. The DD score is a function only of the underlying idea of the response. However, the way in which that idea is expressed is obviously also important in an assessment of primary process. Holt dealt with this by including 47 categories in which to score methods of control or defense manifested in the response. The effectiveness of these defenses is recorded by yet another score, Defense Effectiveness (DE). This rating is a combined measure of a response's form level and the affect accompanying it, with a positive DE score reflecting a well-defended response and a negative score indicating one that was poorly defended.

Role-Taking Task

The assigned content of the RTT revolved around two themes, "hurt" and "please." Hurt was chosen to stimulate stories containing both aggressor and victim, with the role reciprocals viewed as central to the choice of suicide as a symptom behavior. The control stimulus was please, a theme chosen for its relatively neutral quality with regard to social desirability and dynamic impact.

Each subject was shown a white 5 × 8 inch (12.5 × 20.5 cm) index card on which two identical stick figures, separated by the word HURT, were drawn. The subject was directed to use the card as a take-off point for making up a story about two people, one of whom hurts the other. When the initial story was completed, the subject was shown another card identical to the first, except that an arrow was

drawn under the figure on the left. This time the directions were to "make believe that you are one of the characters in the story, the figure on the left. Tell the same story that you told before, but retell it from the perspective of that person, as he or she would tell it." Following completion of this story, the subject was shown a third card identical to the others, except that the arrow was now under the figure on the right. Instructions for this card were the same as for the second card, except this time the subject was asked to take the role of the other character in the story.

The subject was then shown another three-card series, with the word PLEASE replacing the word HURT on otherwise identical cards, and was directed to go through the RTT procedure again, this time telling a story about two people, one of whom does something to please the other. Half of the subjects were presented with the please theme before the hurt theme.

The two sets of stories produced by each subject were then scored to determine the subject's decentering level for each of the two themes. The scores obtained for the please stories were taken as a baseline of the subject's ability to coordinate roles and their reciprocals in a relatively neutral situation. Scores for stories containing the roles of victim and aggressor were used to indicate the change in the subject's cognitive functioning when dealing with defensively isolated aspects of personality.

All stories were scored twice, in accordance with the RTT coordination method and the RTT differentiation method. Although both methods assess the subject's ability to take different role perspectives, they provide different measures of this ability. The RTT coordination method focuses on the degree and quality of coordination between the different perspectives. Increasingly higher scores are given to reflect superior ability to assume different roles and shift from one to the other while maintaining inner-role consistency and balance in the situation described. At one extreme are the categories for obvious inconsistency between the characters' viewpoints. At the other extreme, the subject manifests an ability to synthesize the two perspectives, with each character mindful of the external role orientation of the other.

The RRT differentiation method provides a measure of the subject's ability to differentiate and coordinate attributes of the "self" and "other" within a single retelling of the initial story. Each retelling provides a measure of decentering ability from the perspective of one given role. The subject's ability to decenter was thus measured in four different situations, as victim, as aggressor, as "pleaser," and as "pleased," or in other words, as the subject of dynamically isolated role attributes and as the subject of attributes not dynamically isolated. Low scores indicate little or no shifting of perspective from one character to the other, whereas higher scores reflect an increasingly sophisticated ability of one character to be taken as the object of an internalized state of the other.

Table 1

Comparisons Within Groups of Responses With Libidinal or Aggressive Content

Group	n	Libidinal	Aggressive	T ^a
Mean percentage				
Attempters	19	11.25	23.25	29.5**
Completers	18	8.10	18.05	27.5**
Nonsuicidals	17	13.80	17.60	30.5*
Mean defense demand				
Attempters	19	1.48	2.00	27.5**
Completers	20	1.80	2.22	73.0
Nonsuicidals	18	1.83	2.34	63.0
Mean defense effectiveness				
Attempters		.45	.58	122
"	14		.19	
Completers		1.02	.95	98
"	14		.20	
Nonsuicidals		.68	.70	118
"	15		.18	

^a Wilcoxon matched-pairs signed-ranks test.

^b Mann-Whitney U.

* $p < .05$.

** $p < .005$.

Reliability of Scoring

Protocols from the Rorschach and the RTT were scored by both the examiner and another scorer, a graduate student in clinical psychology. Neither scorer knew to which subject group the protocol being scored belonged. Interjudge reliability coefficients were obtained for scores of the Rorschach and the RTT. The overall interjudge correlation on the Rorschach scores was .83. The coefficients were higher for the RTT. For the coordination scores the reliability coefficient was .88, and it was .96 for the differentiation scores. For all-numerical scores on which the judges differed, the final score was the average of the original two.

Results

The results of this study fall into two groups, those that indicate the subjects' personality functioning as reflected in their Rorschach responses and those based on the RTT scores that describe their cognitive functioning in interpersonal situations. The analyses of the results of this study were done with nonparametric tests, as the results were not normally distributed and did not have equal variances.

Table 2

Comparison of All Rorschach Scores: Suicide Attempters and Completers

Group	<i>M</i> %	<i>n</i>	<i>U</i> ^a	<i>M</i> DD	<i>n</i>	<i>U</i> ^a	<i>M</i> DE	<i>n</i>	<i>U</i> ^a
Aggressive									
Attempters	23.25	20	160	2.00	20	182	.58	19	156
Completers	18.05	20		2.22	20		.95	20	
Libidinal									
Attempters	11.25	20	163	1.48	20	183	.45	14	73
Completers	8.10	20		1.80	20		1.02	14	

Note. DD = defense demand; DE = defense effectiveness.

^a Mann-Whitney U; all values are ns.

Personality Functioning: Rorschach Results

Table 1 summarizes the within-group comparisons for all three subject groups. As predicted, the suicidal subjects, both attempters and completers, experienced breakthroughs of significantly more aggressive than libidinal drive material in responding to the Rorschach. The nonsuicidal control group also had a greater percentage of aggressive than libidinal drive-related responses. However, the level at which that difference is significant was lower for the control group than for either of the suicidal groups.

The aggressive responses of the suicide attempters were significantly more primitive than their libidinal responses as indicated by the differences in DD scores. Comparison of those scores for the nonsuicidal control group did not yield significant differences between their aggressive and libidinal responses. The degree to which their responses were rated as primitive or bizarre was not a function of the particular drive content of the responses, although this was the case for the suicide attempters.

The suicide completer group also showed no significant differences between the DD scores of their aggressive and libidinal responses. This was the one instance in which their results did not parallel those of the attempter group.

Contrary to the prediction, no significant differences were found between the DE scores of the libidinal and aggressive responses for both suicidal groups. It therefore appears that even the attempter group, which experienced

the breakthrough of a great deal of primitive aggressive drive material, defended against their aggressive responses as well as they did against their libidinal responses. The completer group also was equally effective in dealing with the aggressive and libidinal material in the records.

Comparisons between the three subject groups on each of the Rorschach measures studied were all nonsignificant. The suicidal groups did not seem to be dealing with more aggressive drive material than was the nonsuicidal group. In addition, the suicidal group's defenses against those impulses were not more primitive or less effective than were the defenses of the control group. Table 2 summarizes the between-group comparisons for the two suicidal groups, and Table 3 contains the results of comparisons between the suicidal and control groups.

Cognitive Functioning: RTT Results

The results of the RTT coordination scores fully support the hypothesis that suicidal subjects function at a more primitive cognitive level than nonsuicidal subjects when presented with aggressive stimuli. Table 4 summarizes the within-group comparisons of coordination scores. Suicidal subjects were clearly deficient in their ability to maintain a consistent theme while taking the roles of both aggressor and victim, when their performances were compared to their ability to coordinate perspectives in a neutral situation. Nonsuicidal subjects manifested no such difficulty when their perform-

Table 3
Comparison of All Rorschach Scores: Suicidal and Control Groups

Group	M%	n	U ^a	M DD	n	U ^a	M DE	n	U ^a
Aggressive									
Suicide attempters	23.25	20		2.00	20		.58	19	
Nonsuicidals	17.60	20	180	2.34	20	187	.70	18	140
Suicide completers	18.05	20	195	2.22	20	197	.95	20	154
Libidinal									
Suicide attempters	11.25	20		1.48	20		.45	14	
Nonsuicidals	13.80	20	193	1.83	20	182	.68	18	93
Suicide completers	8.10	20	154	1.80	20	194	1.02	14	107

Note. DD = defense demand; DE = defense effectiveness.

^a Mann-Whitney U; all values are *ns*.

ances with the aggressive and neutral themes were compared.

Although the coordination scores of the suicidal subjects for the hurt situation were significantly lower than their scores in the please situation, it is important to note that both of the mean coordination scores of the nonsuicidal control group were lower than either of the suicidal groups. Possibly the control group's lower coordination scores were a manifestation of the effect that their different psychopathologies had on their cognitive performance in that interpersonal situation. Perhaps the suicidal subject, in dealing at an unconscious level with aggressive impulses, has overemphasized the role of benevolent or nurturing person. In any case, this difference between the groups requires that the coordination scores of each subject group in the aggressive situation be compared to their scores in the neutral situation in order to measure the effect of the aggressive stimulus on performance. The mean difference score of the suicidal group differed significantly from that of the control group, with the suicidal subjects generally performing at a higher level in the neutral situation and the nonsuicidal subjects doing better in the aggressive situation. The mean difference between hurt and please for the suicidal and nonsuicidal groups was 2.55 and 1.45 respectively ($U = 132$, $p = .05$).

The results of the analyses of the differentiation scores of the two subject groups also support the hypotheses. The cognitive functioning of the suicidal subjects when taking the isolated role of aggressor was significantly more primitive than their performance in the ego-syntonic role of victim. There were no significant differences in their abilities to assume the role reciprocals with a neutral theme or in the nonsuicidal subjects' performances in any of the four roles. Interestingly, the suicidal subjects achieved their lowest differentiation scores among all roles when taking the role of aggressor and the highest of all their differentiation scores for the role of victim. These results are displayed in Table 5.

Comparisons between the two groups for the differentiation scores, found in Table 6, also support the prediction. The magnitude of the difference between the suicidal group's scores in the roles of aggressor and victim was significantly greater than the magnitude of the difference between the control group's scores for those roles. There were no such findings in the between-group comparison for the neutral situations.

Discussion

This study examined two major areas of functioning, cognitive and defensive. The results fully supported the hypotheses about

Table 4
*Comparisons Within Groups of Role-Taking
 Task Coordination Scores for Dynamically
 Isolated and Relatively Neutral Role
 Reciprocals*

Group	M coordination		T ^a
	Hurt	Pleased	
Suicidal	18.15	20.90	37*
Nonsuicidal	17.95	16.50	70

Note. $n = 18$.

* Wilcoxon matched-pairs signed-ranks test.

* $p < .025$.

cognitive functioning. The suicidal subjects differed from the control group in a very defined manner. When asked to take the reciprocal roles of aggressor and victim in a story about one person hurting another, a situation that highlighted what is presumably the central conflict in suicidal behavior, these subjects had difficulty in coordinating the same thematic material from two opposite perspectives. In addition, they were more adept at assuming the ego-syntonic role of victim than the ego-alien role of aggressor, with several of the suicidal subjects actually commenting that they "just couldn't" tell the story from the aggressor's point of view. These same subjects performed significantly better on all measures when taking roles in the dynamically neutral situation of one person pleasing another. They had less difficulty coordinating the two roles and took the roles of both

Table 5
*Comparisons Within Groups of Role-Taking
 Task Differentiation Scores for Dynamically
 Isolated and Relatively Neutral Roles*

Group	M differentiation		T ^a
	Aggressor	Victim	
Suicidal (17)	2.67	3.33	21.0*
Nonsuicidal (14)	2.75	2.44	33.5
	Pleaser	Pleased	
Suicidal (17)	3.27	3.14	80.5
Nonsuicidal (15)	2.68	2.85	52.0

Note. Numbers in parentheses are n s.

* Wilcoxon matched-pairs signed-ranks test.

* $p < .005$.

pleaser and pleased with equal facility. For the nonsuicidal subjects, there were no significant differences in their performance in either the aggressive or neutral situation, and their scores in the aggressive situation were significantly higher than those of the suicidal group.

The results of this study that pertain to defensive functioning partially supported the hypotheses. As predicted, suicidal subjects experienced breakthroughs into consciousness of more aggressive drive-related than libidinal drive-related material. Subjects who later committed suicide also seemed to have been dealing with more aggressive impulses during testing. In addition, the aggressive responses of subjects who had attempted suicide were more primitive and more bizarre than their libidinal responses.

Some of the results did not turn out as predicted. Although, as noted above, the aggressive responses of the suicide attempters were more primitive than their libidinal responses, there were no significant differences in the suicidal subjects' ability to defend against either type of response. There would seem to be at least two explanations for this. It may be that Holt's DE score, used in this study as a measure of the sophistication of a subject's defensive functioning, was not the most appropriate instrument for the task. That is, individuals may defend against bizarre or primitive responses in ways that are more or less immature or pathological. However, if the defense mechanisms, regardless of their sophistication, succeed in protecting the subjects from conscious anxiety or confusion, then they have been effective and would be scored as

Table 6
*Comparisons of Mean Differences Between
 Differentiation Scores for Dynamically Isolated
 and Relatively Neutral Role Reciprocals*

M difference between	Suicidal	Nonsuicidal	U ^a
Aggressor-victim	-.67	+.36	92*
Pleaser-pleased	+.06	-.31	194

Note. $n = 20$ for all four conditions.

* Mann-Whitney U.

* $p < .01$.

such. So, in interpreting these findings, it could be said of suicidal individuals that though in aggressive situations they may indeed use defenses that are developmentally immature, they do not necessarily decompensate in such cases to any greater degree than in other circumstances. Indeed, nowhere in the literature is it suggested that a clear sign of high suicide potential is a person's overt disorganization in response to aggression. Rather, the response is said to be an unconscious regression of the ego to more primitive forms of defense.

There is another explanation for the lack of significant differences between the libidinal and aggressive DE scores of the suicidal subjects. It is possible that there actually was no difference in the maturity of these subjects' defensive responses to the breakthrough of various types of drive-related material. Their ego functioning was not clearly defined under a variety of different circumstances, instead it remained fairly constant throughout. Such an interpretation can be expanded to deal with the other two findings in this study that were not predicted by the hypotheses.

It has been expected that significant differences between suicidal and nonsuicidal subjects would appear in their personality functioning in aggressive situations. This was not the case. There were no significant differences between the three groups on any of the Rorschach scores. Furthermore, just as with the two suicidal groups, the nonsuicidal psychiatric controls also experienced the breakthrough into consciousness of more aggressive than libidinal material on the Rorschach, although the degree of difference was not nearly as high for the control group. These results may be used to bring into question the psychoanalytic theory of the significance of aggression in suicide.

However, the similarity between the suicidal and nonsuicidal psychiatric groups can also be understood as a reflection of the importance of aggressive conflicts not only in self-destructive behavior but also in a number of other diagnostic entities. Jacobson (1971), in attempting to combine ego psychology with psychoanalytic drive theory, implicated aggression as a major factor in all

depression, whether suicidal ideation is present or not. Silverman (1966), using the technique of subliminal stimulation, showed that unconscious aggressive impulses are directly related to ego pathology in schizophrenia.

One therefore might well view various forms of psychopathology not as discrete entities but as forming a continuum of symptomatic expressions of universal conflicts. Both suicidal and nonsuicidal psychiatric subjects must deal with a large amount of aggressive drive-related material, because aggressive conflicts are a major factor in pathological ego functioning. Perhaps the overall severity of the pathology, whatever its form, is a function of the extent to which drive-related material breaks into consciousness.

This is not to imply that any such breakthrough is pathological. It is clear that aggressive impulses are an important component of the healthiest personality. The presence of these impulses can be seen as evidence of ego dysfunction only when it leads to severe anxiety or when aggression dominates the personality to the exclusion of other drives. Therefore, if so-called normal control subjects were tested and compared with suicidal or other hospitalized subjects, differences might be found in the percentages of aggressive responses in the total record or in the predominance of aggressive over libidinal material. But one could not expect to find such differences when dealing exclusively with a hospitalized population.

For the same reason, no significant differences were found either within or between groups for DE scores. As discussed above, these scores can be understood as measures of the individual's overall ego functioning, and none of the three groups functioned significantly more poorly than any other.

If aggression plays a major role in both schizophrenia and depression, whether suicidal ideation is present or not, what dictates the form in which the pathology will manifest itself? What determines the various symptomatic expressions of general ego dysfunction?

The results of this study suggest that the individual's cognitive functioning, the area in which the clearest differences between the suicidal and nonsuicidal groups were found, may

define the environment and thus determine the nature of symptom expression. Suicidal subjects do not differ significantly from their nonsuicidal psychiatric counterparts in overall level of ego pathology or overall level of cognitive functioning. However, they do show more markedly primitive functioning in interpersonal situations with aggressive content. In this area there is evidence of clear isolation among suicidal subjects between the role reciprocals of aggressor and victim. They are at their best cognitively when taking the role of victim, but they have so dissociated themselves from the ego-alien role of aggressor that perspective in the aggressive situation remains relatively unavailable to them.

One could say that they have thus defined their environment as one that precludes the direct outward expression of aggression, a relatively stable condition that in itself might not give rise to pathological behavior. However, when external events lead to the conscious experience of aggressive drive derivatives, the suicidal subjects can deal with the resulting breakdown of ego defenses only by treating themselves as victims. At deeper levels of the unconscious, such actions may indeed succeed in killing introjected others, but consciously the suicidal subjects can perceive the situation only as one in which they alone have been hurt. Because nonsuicidal subjects are able to assume the role of aggressor in an aggressive situation, self-destructive behavior is a less likely, and certainly a less necessary, outcome of the breakthrough of aggressive drive derivatives, though it may lead them to ego pathology as serious as that of the suicidal group.

Reference Note

1. Holt, R. R. *Manual for scoring primary process material in the Rorschach*. Unpublished paper, New York University, 1970. (Mimeo)

References

- Eisenthal, S. Suicide and aggression. *Psychological Reports*, 1967, 21, 745-751.
- Feffer, M. The cognitive implications of role-taking behavior. *Journal of Personality*, 1959, 27, 152-167.
- Fisher, S. The value of the Rorschach for detecting suicidal trends. *Journal of Projective Techniques*, 1951, 15, 250-254.
- Freud, S. Mourning and melancholia. In, *Standard edition of the complete psychological works*. London: Hogarth Press, 1957. (Originally published 1917)
- Freud, S. *Beyond the pleasure principle*. New York: Liveright, 1950. (Originally published 1920)
- Jacobson, E. *Depression: Comparative studies of normal, neurotic and psychotic conditions*. New York: International Universities Press, 1971.
- Lester, D. Resentment and dependency in the suicidal individual. *Journal of General Psychology*, 1969, 81, 137-145.
- Lowenherz, L., & Feffer, M. Cognitive level as a function of defensive isolation. *Journal of Abnormal Psychology*, 1969, 74, 352-357.
- Menninger, K. *Man against himself*. New York: Harcourt, Brace & World, 1938.
- Piaget, J. *The psychology of intelligence*. New York: Harcourt, Brace, 1950.
- Raphling, D. L. Dreams and suicide attempts. *Journal of Nervous and Mental Disease*, 1970, 151, 404-410.
- Silverman, L. H. A technique for the study of psychodynamic relationships: The effects of subliminally presented aggressive stimuli on the production of pathological thinking in a schizophrenic population. *Journal of Consulting Psychology*, 1966, 30, 103-111.
- Vinoda, K. S. Personality characteristics of attempted suicides. *British Journal of Psychiatry*, 1966, 112, 1143-1150.
- Ward, C. Delusions as a form of primitive decentering. (Doctoral dissertation, Yeshiva University, 1975). *Dissertation Abstracts International*, 1976, 36, 5823B. (University Microfilms No. 76-10326)
- Zilboorg, G. Considerations on suicide, with particular reference to that of the young. *American Journal of Orthopsychiatry*, 1937, 7, 15-31.

Received March 9, 1977 ■

Relationships Between WISC-R Factors, Wide-Range Achievement Test Scores, and Visual-Motor Maturation in Children Referred for Psychological Evaluation

James M. Stedman

Department of Psychiatry
University of Texas Health Science
Center at San Antonio and
Community Guidance Center, San Antonio

Robert H. Cortner

University of Texas Health Science
Center at San Antonio and
Community Guidance Center, San Antonio

G. Frank Lawlis

North Texas State University

Gloria Achterberg

Community Guidance Center, San Antonio

The present study investigated relationships between the Kaufman Wechsler Intelligence Scale for Children-Revised factors (Verbal Comprehension, Perceptual Organization, and Freedom from Distractability), Wide-Range Achievement Test (WRAT) scores, and visual-motor maturation in a sample of 106 children. These children ranged in age from 6 to 13, and they had been referred for clinical evaluation because of a variety of school-related problems, including learning and classroom behavior problems. Results indicated significant correlations between the Verbal Comprehension factor and the Reading, Spelling, and Arithmetic measures on the WRAT and all measures of visual-motor maturation. The Perceptual Organization factor correlated significantly with Reading, Spelling, and all visual-motor maturation measures. The Freedom from Distractability factor correlated significantly with Arithmetic.

Since Luty (1967) and Sattler (1974) devised methods for the clinical use of Cohen's (1959) factor analysis of the Wechsler Intelligence Scale for Children (WISC), clinicians have increasingly used a factor score approach of WISC interpretation. No doubt this tradition will continue with Kaufman's (1975) factor analysis of the Wechsler Intelligence Scale for Children-Revised (WISC-R); and, in fact, Sattler has suggested WISC-R interpretations. Given this increasing trend toward use of the factor approach in clinical interpretation, it would seem important to establish correlational networks between WISC-R factor scores and other criteria, par-

ticularly those tests most commonly used by clinicians.

The present study sought to investigate relationships between the Kaufman factors Verbal Comprehension (VC), Perceptual Organization (PO), and Freedom from Distractability (FD); achievement, as reflected on the Wide-Range Achievement Test (WRAT; Jastak & Jastak, 1965); and perceptual maturation, as reflected in the Koppitz (1964) scoring of the Bender-Gestalt in a sample of school children referred for psychological evaluation. The following relationships were hypothesized: (a) Since the VC and PO factors, as identified by Kaufman, are equivalent to the Verbal and Performance IQ scores on the WISC-R, a positive relationship with Reading and Spelling on the WRAT was anticipated. (b) Since Verbal IQ has been shown to be positively related to WRAT Math scores (Jastak & Jastak, 1965) and since math is thought to be sensitive to dis-

Gloria Achterberg is currently a student at North Texas State University.

Requests for reprints should be sent to James M. Stedman, Department of Psychiatry, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, Texas.

Table 1

Intercorrelation Matrix of Wechsler Intelligence Scale for Children - Revised Subtests

Test	1	2	3	4	5	6	7	8	9	10
1. Information										
2. Similarities	.58									
3. Arithmetic	.57	.52								
4. Vocabulary	.68	.60	.56							
5. Comprehension	.68	.59	.51	.70						
6. Picture Completion	.21	.31	.31	.39	.31					
7. Picture Arrangement	.38	.43	.34	.44	.39	.29				
8. Block Design	.31	.32	.26	.42	.27	.44	.31			
9. Object Assembly	.38	.36	.35	.38	.42	.53	.30	.43		
10. Coding	.23	.14	.36	.16	.18	.10	.21	.11	.36	
11. Digit Span	.36	.33	.34	.36	.36	.22	.36	.33	.21	.27

tractability, it was anticipated that the FD and the VC factors would be correlated with Arithmetic achievement on the WRAT. (c) It was anticipated that the PO factor would be positively correlated with Koppitz scores on the Bender-Gestalt.

Method

Subjects

The subjects for this study were students (76 males and 30 females). These students ranged in age from 6 to 13 with a mean age of 9.5 and were distributed across grade levels as follows: first, 13.2%; second, 16.9%; third, 17.9%; fourth, 11.3%; fifth, 10.4%; sixth, 14.1%; seventh, 7.5%; and eighth, 8.4%. Their IQs ranged from 60 to 118 with a mean of 88.1 and a standard deviation of 13.1. All were referred for evaluation because of learning and/or classroom behavior problems. Of these 106 subjects, 90% had a Spanish surname, and it is likely that a high percentage of these subjects are bilingual. The remainder were Anglo or black, with blacks constituting only 2% of the total sample.

Procedure

As part of a service contract with the parochial school system, Archdiocese of San Antonio, WISC-R, WRAT, and Bender protocols were collected during the 1975-1976 school year. Because of a desire to study factor structure per se, the WISC-R protocols were intercorrelated with respect to the 11 subtests, resulting in an 11 × 11 matrix. (The Maze subtest was not administered.) A principal-components analysis (with squared multiple correlations in the diagonal) was performed, terminating with an eigenvalue less than 1. Three factors, accounting for 64.6% of the variance, were retained and rotated according to the varimax criteria. Factor scores were computed for each subject, and these scores were transformed into normal distribution.

These factor scores, which were essentially equivalent to the Kaufman factors, were then correlated with age-appropriate WRAT Level 1 or Level 2 standard scores for Reading, Spelling, and Arithmetic and Koppitz error scores for the Bender-Gestalt protocols. Because Koppitz raw scores are more normally distributed in the 6- to 10-year-old age range, correlations were calculated only with subjects up through the fifth grade ($n = 77$). These error scores included total number of scorable errors, total significant, and total highly significant errors. Significant and highly significant errors are those that Koppitz (1964) found to occur in protocols of children diagnosed as brain injured.

Results

Table 1 presents the intercorrelation matrix for the 11 subtests, and Table 2 presents the factor structure obtained by the methods described above.

Table 3 presents Pearson correlations between the three factor scores and the WRAT standard scores for Reading, Spelling, and Arithmetic. As anticipated, the VC factor contributed significantly to Reading scores ($p < .005$), but it should be noted that the PO factor also was significantly related ($p < .05$), although the magnitude of the relationship was considerably less. The VC factor also was significantly related ($p < .005$) to Spelling scores, and again the PO factor made a small but significant contribution ($p < .05$). The FD factor, as expected, related significantly ($p < .005$) to the Arithmetic score and, as anticipated, the VC factor also contributed ($p < .05$). The PO factor apparently contributes little to Arithmetic as measured by the WRAT.

Table 4 gives Pearson correlations between the three factor scores and the total Koppitz Bender errors, the total significant errors, and the total highly significant errors. As predicted, there was a significant relationship, expressed negatively, between the PO factor and absolute and highly significant Koppitz Bender scores ($p < .05$ to $p < .005$). Of note are the significant ($p < .01$ to $p < .005$) associations of the VC factor with all Koppitz scores and the significant ($p < .01$ to $p < .005$) associations of the FD factor with both the absolute and significant Koppitz scores.

Discussion

The present study confirms that the Kaufman factors are related in clinically expected ways to WRAT achievement scores and visual-motor maturation scores in a sample of children referred for psychological evaluation. This finding is perhaps reassuring to

Table 2
Factor Structure of the Wechsler Intelligence Scale for Children-Revised

Test	Unrotated first factor	Varimax rotation		
		VC	PO	FD
Verbal				
Information	.78	.84	.10	.16
Similarities	.75	.77	.23	.04
Arithmetic	.72	.66	.13	.40
Vocabulary	.82	.82	.30	.03
Comprehension	.78	.82	.18	.09
Digit Span	.55	.38	.23	.42
Performance				
Picture Completion	.55	.16	.83	.01
Picture Arrangement	.61	.48	.31	.21
Block Design	.56	.23	.75	.04
Object Assembly	.64	.23	.68	.36
Coding	.38	.07	.05	.94

Note. VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom from Distractibility.

Table 3
Correlations Between WISC-R Factors and WRAT Standard Scores

WISC-R factor	WRAT		
	Reading	Spelling	Arithmetic
VC	.44**	.38**	.20*
PO	.19*	.17*	.10
FD	.04	.14	.37**

Note. WISC-R = Wechsler Intelligence Scale for Children-Revised; WRAT = Wide-Range Achievement Test; VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom from Distractibility.

* $p < .05$.

** $p < .005$.

clinicians who, all too often, see Bender protocols that do not seem to fit at all with findings on the PO factor and distracted, anxious children with good Arithmetic and Digit Span scores.

Other relationships are also of interest. The PO factor's small but significant contribution to Reading and Spelling indicates that nonverbal intellectual factors make their contribution to those skills, as well as to Arithmetic. Also, both the VC factor and the FD factor were significant, suggesting that verbal mediators and attention factors are involved in visual-motor maturation processes.

Table 4
Correlations Between WISC-R Factors and Koppitz Bender-Gestalt Errors

Factor	No.	Total errors	
		Significant	Highly significant
VC	-.43***	-.27**	-.41***
PO	-.30***	-.15	-.22*
FD	-.28**	-.44***	.08

Note. WISC-R = Wechsler Intelligence Scale for Children-Revised; VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom from Distractibility.

* $p < .05$.

** $p < .01$.

*** $p < .005$.

Perhaps, all of these additional relationships are not surprising, but they do suggest areas for further research regarding relative WISC-R factor contributions to related achievement and visual-motor maturation measures.

Finally, it should be noted that these data suggest that perhaps the factor structure for bilingual Mexican Americans is approximately equivalent to that of English-speaking subjects. Although this conclusion cannot be stated definitely due to the mixed nature of the sample, the evidence seems to point in that direction. Further research is also needed to ascertain whether bilingual factor structures are indeed equivalent.

References

- Cohen, J. The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *Journal of Consulting Psychology*, 1959, 23, 285-299.
- Jastak, J. F., & Jastak, S. R. *The Wide Range Achievement Test*. Wilmington, Del.: Guidance Associates, 1965.
- Kaufman, A. S. Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology*, 1975, 43, 135-147.
- Koppitz, E. M. *The Bender-Gestalt Test for young children*. New York: Grune & Stratton, 1964.
- Luty, C. *Individual intelligence tests: A manual*. Greeley, Colo.: Author, 1967.
- Sattler, J. M. *Assessment of children's intelligence*. Philadelphia: Saunders, 1974.

Received March 14, 1977 ■

INSTRUCTIONS TO AUTHORS

Style of manuscripts. Manuscripts must be prepared in the style described in the *Publication Manual of the American Psychological Association* (2nd ed.). Instruction on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors should submit manuscripts in triplicate and should keep a copy to guard against loss.

Abstracts. Manuscripts of regular articles must be accompanied by an abstract of 100-175 words. Manuscripts of brief reports must be accompanied by an abstract of 75-100 words. All abstracts must be typed on a separate sheet of paper.

Brief Reports. The *Journal of Consulting and Clinical Psychology* will accept Brief Reports of search studies in clinical psychology. The procedure is intended to permit the publication of soundly designed studies of specialized interest or limited importance which cannot now be accepted as regular articles because of lack of space. Several pages in each issue will be devoted to Brief Reports, published in the order of their receipt without respect to the dates of receipt of the regular articles. Most Brief Reports appear in the first or second issue to go to press following their final acceptance.

An author who wishes to submit a Brief Report:

1. Sends the Brief Report, limited to two printed pages and prepared according to the specifications given below.

2. Agrees not to submit the full report to another journal of general circulation.

Specifications. The Brief Report should give a clear, condensed summary of the procedure of the study and as full an account of the results as space permits.

To ensure that a brief report does not exceed two printed pages, follow these instructions for typing: (a) Set typewriter to a 48-space line. (b) Type text. (c) Count all lines except abstracts, title, and by-line, including acknowledgments. If you have exceeded 185 lines, shorten the material. (d) If your brief report barely exceeds 75 lines (one printed page), try to edit to 75 lines exactly. Reports slightly exceeding one page will be edited for length in order to use space economically.

In Brief Reports, headings, tables, and references are avoided or, if essential, must be counted in the 75 lines. This journal no longer requires an extended report, but, if one is available, each Brief Report must be accompanied by the following footnote:

Requests for reprints and for an extended report of this study should be sent to John Doe (give the author's full name and address).
The footnote should be typed on a separate sheet and not counted in the 75-line quota.

Freedom of Choice and Behavioral Change

Frederick H. Kanfer and Laurence G. Grimm
University of Illinois

A 3×2 (Treatment \times Population) factorial design with repeated measures (pretest/posttest) was used to evaluate the effects of perceived freedom of choice on behavior change in a therapy analogue study. Ninety subjects were assigned to three groups that varied in the amount of perceived choice given to subjects in determining the type of training procedure used for speed-reading enhancement. Experimental conditions were crossed with two populations of subjects to examine two levels of perceived freedom. Half of the subjects were drawn from a pool of psychology students required to participate in psychology experiments, and the remaining half of the subjects were volunteers. The main dependent variable was the amount of change in reading rate. A marginally significant ($p = .06$) increase in reading speed was obtained by volunteer subjects in comparison to subject pool participants. Subjects who perceived that they were given a choice in training procedures improved significantly more ($p < .02$) in reading speed than subjects who lost the freedom of choice. No changes in reading comprehension were noted. These findings are discussed in terms of the relationship between freedom of choice and performance in a behavior change program.

Increased sophistication about psychological treatment, a social trend toward self-help therapies, and the declining doctrine of the infallibility of mental health professionals have resulted in greater consumer selectivity among therapy agents and methods. With greater client understanding of behavior change methods, the client's acceptance of a treatment program has become even more critical for therapeutic improvement than before. It has long been established that motivation to change is essential for treatment success. Clinical folklore and current psychological theories suggest ample reasons for enhancing a client's belief that the treatment is voluntary and that he or she has some voice in deciding among treatment ap-

proaches in order to further cooperation. In fact, in many cases, especially when referred by agencies or other persons, the development of a motivation to change becomes a first treatment objective (Kanfer, 1975).

Laboratory findings closely match clinical experience. Liem (1975) found that students who were given a choice in selecting among different types of classroom sections did significantly better on exams and reported greater satisfaction with their sections and class leaders in comparison to no-choice subjects. Similarly, residents in a home for the aged who were given greater personal responsibility and choice in their daily routines showed significant improvement in alertness, active participation, and self-rated well-being than control subjects (Langer & Rodin, 1976). The belief that one has a choice also appears related to perception of increased control over an outcome. For example, choice of the order of taking a number of tests can reduce anxiety (Stotland & Blumenthal, 1964). The presence of an escape response from an aversive noise decreased the aversiveness of the threatening stimulus (Corah &

This study was supported in part by a grant from the University of Illinois Research Board.

The authors wish to thank Donald L. Barth and Robert Cherney for serving as experimenters.

Laurence G. Grimm is now at the University of Illinois at Chicago Circle.

Requests for reprints should be sent to F. H. Kanfer, Psychological Clinic, University of Illinois, 505 East Green Street, Champaign, Illinois 61820.

Boffa, 1970). Thus, a belief in personal control strongly influences the tolerance of aversive events (Averill, 1973; Kanfer & Seidner, 1973; Langer, Janis, & Wolfer, 1975). Langer (1975) found that choice in a chance situation increased confidence and risk taking. The freedom to choose among alternatives is thus related to perception of personal control.

The clinical implications of attributing a behavior change to one's own actions lie in the findings of better maintenance of that behavior (Davison & Valins, 1969; Kopel & Arkowitz, 1975). The finding of the facilitating effects of freedom of choice is supplemented by the finding that elimination of freedom or a threat to eliminate freedom has the opposite effects in activating the person to oppose such restrictions (Brehm, 1966). For example, a study by Brehm and Brehm (see Brehm, 1966) showed that subjects resisted persuasion when an authority figure acted in a dictatorial manner (i.e., "You must agree.") When the communication was phrased so as to not pose a threat to freedom, it was highly persuasive. Experiments on the effect of reactance suggest that a threat to eliminate freedom or an intentional deprivation of freedom yields detrimental behavioral changes with regard to an assigned task. The clinical parallel would be the referral of a client for therapy under conditions of coercion. Reactance and countercontrol (Davison, 1973) would be expected to interfere with the treatment process.

The forced referral to the therapist is somewhat paralleled in the experimental situation in which students in psychology courses are required to participate in research. Both the ethical and technical problems of using such subjects have been repeatedly discussed (Kelman, 1972; Schultz, 1969). Cox and Sippelle (1971) demonstrated that "true volunteers" markedly differed from nonvolunteers in mean heart rate changes over trials during an operant verbal conditioning procedure. In a recent study, Gordon (1976) examined the behavior of volunteer and nonvolunteer subjects who could either choose or were assigned to different relaxation treatments. Gordon found that volunteers valued the treatment more and reported it to be significantly more effective than subjects who had no

choice. These findings are also consistent with the work presented by Rosenthal and Rosnow (1975) that emphasized that the use of volunteer versus nonvolunteer subjects introduces important sources of bias in the experimental results.

The present study used a laboratory analogue to examine two critical problems in clinical practice: the client's freedom of choice in deciding to submit to therapy and the client's perceived choice in selecting among therapeutic procedures. A free-choice condition in selecting training procedures was used as an analogue to treatments that emphasize negotiation in establishing change programs. A lost-choice condition represented a situation in which a client enters therapy with expectations for negotiation but is then given the treatment that the therapist recommends. The analogue for this group is somewhat weaker, since in this study the experimenter also manipulated the prior expectations. A no-choice condition corresponded to the client who has no strong expectations about participating in deciding on treatment plans and does not choose a treatment. To increase similarity to clinical procedures, skill in reading was chosen as an area in which college students frequently experience problems. Specifically, this study examined the effect of (a) volunteer versus nonvolunteer participation and (b) choice versus imposition in selection of training procedures on improvement in brief speed-reading practice.

Method

Subjects

The subjects were 90 students at the University of Illinois at Urbana-Champaign. Forty-five subjects (23 males and 22 females) were enrolled in introductory psychology courses in which research participation credit was given toward their course grades. Subjects were required to submit their names for participation and were then assigned by computer to five experiments. The subjects had no choice in this assignment, but they could refuse to participate in a given experiment. The remaining 45 subjects (27 males and 18 females) were volunteers who answered campuswide advertisements that read "Wanted: Participants for a study concerning increased reading speed and comprehension." The nonvolunteers were assigned to a room and were scheduled for experimental participation by routine psychology department procedures. Volunteers con-

tacted the experimenter by telephone to arrange their appointments. Neither the advertising nor initial contact included an explicit or implicit statement about the subjects' opportunity to obtain remedial training. All subjects were randomly assigned to one of three groups in a 3×2 factorial design. Test and questionnaire data indicated that all groups were equivalent in pretest reading speed and comprehension, expectations for change, prior training experience, and motivation to learn speed reading.

Procedure

Subjects were seen individually and were told that the aim of the study was to examine the immediate effects of several different training techniques on reading rate and on comprehension. The long-term purpose of the experiment was said to assist in the development of reading skill programs. All subjects filled out a questionnaire to assess their interest in improving reading skills while the experimenter was seated behind them. After completion of the questionnaire, the subjects were given instructions for the pretest. The pretest consisted of a narrative passage at the low college level used by the University Counseling Center for establishing reading ability. The selected passages were from the book *Toward Better Reading Skills*. The subjects were told to read the passage quickly but to try also to understand the material, since they would be asked questions afterwards. When individual subjects indicated their readiness, the experimenter started a stopwatch and signaled them to begin reading. The amount of time taken to complete the reading of the standard passage was used as a reading speed measure.

After subjects had finished the reading task, they were given a comprehension test, comprised of multiple-choice questions. Following the administration of the comprehension test, subjects were seated in front of a small screen, which was used for reading skill training. Other training materials included a Carousel projector, set up for rear projection, and 50 slides of phrases ranging from 4 to 5 words each. The phrases were taken from a film strip produced by the Society for Visual Education, Inc., Chicago, Illinois, designed to improve reading skills. Each phrase was set up tachistoscopically so that it appeared on the screen for approximately .14 sec. Timers provided an interval of 5 sec between slide presentations. Immediately prior to the presentation of training materials, the subjects were told:

This study is examining training techniques and their effect on reading rate and comprehension. With the techniques that I will describe to you in a moment, it has been shown that even in one training session some improvement is possible. In the long run the techniques are all equally effective, but, due to individual differences, the method the person feels most comfortable with will probably be the most effective because it best fits his/her style of learning.

The remainder of the instructions were varied according to the group to which the subject was assigned. Subjects in the free-choice conditions were told: "We have found that almost everyone prefers to choose the techniques they work with. How about you? Would you also prefer to have a choice of which reading technique to use?" Subjects in the lost-choice group were given the identical instructions. For no-choice subjects the preceding instructions were omitted. All subjects, with the exception of one male and one female, indicated that they would like to have a choice. Since they did not wish to make a choice, these two subjects were excluded from further participation in the study.

Following these instructions, three training techniques were described to subjects in the following manner:

The first technique that can be used in training is a derivation of the Flourens-Derjeu ocular scan method. Although it has been shown to induce some eye strain in the learner, it is learned quickly and offers rapid improvement. The second technique is called the Mueller peripheral scope method. This method has been shown to produce a small amount of eye strain, takes longer to learn, and results in slower improvement. The third technique is the Morgan lateral field method. It involves a moderate amount of eye strain and requires a medium length of time to learn but yields a moderate and immediate improvement in reading.

Following the characterization of the training techniques, the free-choice group was asked, "Now that you have heard the description of the characteristics of each technique, which one would you prefer to work with?" After these subjects made their choice, the experimenter said, "OK, that's fine, we'll use that one." The lost-choice group was asked, "Now that you have heard a description of the characteristics of each technique, which one would you prefer to work with?" After these subjects had stated a preference, the experimenter went to the back of the laboratory, turned on the equipment, returned, and said, "I am sorry but I cannot give you that one, I will have to give you the _____ method." No-choice subjects were told that although these were the three methods generally used in research of this type, this experiment only used the _____ method. In spite of the fact that specific methods were identified, all subjects were exposed to an identical procedure. Prior to the presentation of the slides, all subjects were told:

You will be shown a slide for a fraction of a second, then there will be a brief rest, followed by another slide. The sequence will continue until all slides have been shown. Now, each slide will have a group of words on it. Try to take in the entire line in one look instead of reading across the line from left to right. I will let you know when all the slides have been shown.

The experimenter then went behind a partition and began to show the slides. After all 50 slides had been presented, the experimenter returned and asked the subject to move to the testing table. A second reading and comprehension test, matched for reading difficulty to the pretraining material, was then administered. This administration followed the same procedures outlined for the pretraining test. It should be noted, however, that in the presentation of the pretests and posttests, the passage contents were alternated to control for an order effect. Following the reading of posttest materials, a comprehension test was again given. All subjects were then told that it would require considerable practice over a period of weeks to actually improve their reading skills. However, the experimenter offered to give them the results of their reading tests as well as those of the study and to discuss with them any desire to take a remedial reading course. The subjects were then to fill out an "experimenter evaluation form." The evaluation form contained seven questions, rated from strongly agree (1) to strongly disagree (5). The items contained statements about the experimenter's manner, his competency, and his presentation of the material. After completion of the experimenter evaluation form, subjects were asked to deposit the form in the second author's mailbox on a different floor in the same building, immediately after the experiment. Those subjects who requested information about improving their reading skills were given prepared materials referring them to the University Counseling Center where such classes are offered.

Overall Design

The experiment consisted of a factorial design with two subject classifications (volunteers and nonvolunteers) and three treatment classifications (free choice, lost choice, and no choice). For all subjects, pretraining and posttraining scores were obtained for reading speed and comprehension. In addition, the pretraining motivation questionnaire and the posttraining experimenter evaluation questionnaire provided data to ascertain initial level of motivation for improving reading skills and postexperimental reactions to the experiment and the experimenter (as a test of potential reactance).

Results and Discussion

The main measure of the effects of the volunteer and choice variables was the change in reading rates following the 50-item training presentation. Table 1 presents the reading rates of all groups on the prereading and postreading tests and mean change scores. To assess the initial reading rate for all groups, a one-way analysis of variance on pretraining reading scores was performed. The nonsignificant $F(5, 84)$ of 1.67 indicates that the groups did not differ prior to the training sessions.

Table 1
Mean Pretest, Posttest, and Change Scores in Reading Rate

Choice	Nonvolunteers	Volunteers
Free		
Pre	298	346
Post	303	376
Change	5	30 _{a,b,c}
Lost		
Pre	322	287
Post	303	285
Change	-19 _a	-2 _b
No		
Pre	294	336
Post	289	341
Change	-5 _c	5

Note. Significant comparisons at $p < .05$ were obtained for pairs of groups identified by the same subscript. Reading rate was measured in words per minute.

To evaluate the relative change in reading rates among groups, a two-way analysis of variance (Groups \times Populations) was performed on raw change scores in reading rates from the pretest to posttest (Overall & Woodward, 1975, 1976). The main effect for the choice conditions was significant, $F(2, 84) = 3.13$, $p < .05$. A marginally significant main effect was obtained for the volunteer versus nonvolunteer comparison, $F(1, 84) = 3.38$, $p = .06$. Volunteer subjects showed a greater increase in reading rate in comparison to subjects drawn from the introductory psychology subject pool. No significant interaction effect was obtained.

The direction of the significant effect of choice conditions is indicated by the array of mean change scores. Subjects in the free-choice condition showed the greatest change, followed by the no-choice and lost-choice subjects. A Scheffé test was performed to locate the sources of the differences demonstrated by the significant main effect. The free-choice groups showed a significantly greater increase in reading rate relative to the lost-choice groups ($p < .02$). These results indicate that the opportunity to choose among alternative training techniques significantly affected performance as a result of the training session.

A two-way analysis of variance (Groups \times Population) on raw gain scores was used to assess changes in reading comprehension scores. The results indicated that neither volunteering nor choice conditions significantly affected reading comprehension changes. Reading rate and comprehension change scores were then transformed to residual change scores and were intercorrelated (Tucker, Damarin, & Messick, 1966). The correlation coefficient was nonsignificant ($r_{xy} = .06$). These findings indicate that changes in reading rates occurred independently of comprehension, and thus faster reading did not lower comprehension.

Since both male and female subjects and two experimenters participated, a two-way analysis of variance was used to test for the effects due to the experimenters as well as for sex. Neither analysis yielded a significant result.

The overall results are consistent with findings in other studies that subjects who perceived that they had a choice among alternatives performed more effectively than subjects who lost such a choice. The implications of these findings for clinical procedures are clear. Treatments that emphasize negotiation about therapy objectives and use participation of the client in the treatment process (Kanfer, 1975) would be expected to work more efficiently toward the therapeutic goals than clients on whom these goals are imposed or for whom the parameters of a procedure are not discussed. Champlin and Karoly (1975) have reported a preliminary investigation in which clients who participated in negotiating contract objectives showed significantly greater activity in the change program than those on whom contract conditions were imposed.

The present data do not permit a clear inference as to whether this difference was due to reactance (i.e., a decline in reading rate) in the lost-choice subjects or an enhancement (a positive motivational effect) in the free-choice group. Failure to obtain statistical differences between the no-choice group and the lost-choice group by the Scheffé test permits no firm conclusion in this matter. An analysis of within-group changes over trials also fails to clarify this issue, since only the nonvolunteer lost-choice group

showed a significant change, that is, a decrease in reading rate ($p < .03$). However, inspection of the data suggests that both effects may have taken place, possibly interacting with the characteristic of the subject population. In contrast to the average increases in reading rates among subjects in the free-choice groups, both volunteer and nonvolunteer subjects in the lost-choice groups showed an overall decrease in reading rate on the posttraining test. In the no-choice group, the rate of volunteers tended to increase and the rate of nonvolunteers tended to decrease after the training session. Variation among subjects and the size of the sample may have contributed to the failure to clarify this question in the present experiment.

The difference in change between volunteer and nonvolunteer subjects just missed significance at the traditional .05 level. However, the strong trend ($p = .06$) supports suggestions from earlier studies and extends the findings by Gordon (1976) that persons who seek out a particular change experience appear to benefit more from it than those who are assigned to it by others. Thus, the implications for clinical practice lie in the need for improvement of clinical techniques that would help a person become involved in a change program, even when initially referred by someone else.

The marginally significant difference between volunteers and nonvolunteers could be viewed as a reflection of the differences in experience with or enthusiasm about a reading skills program. The questionnaire data indicated that 31% of the nonvolunteer and 20% of the volunteers had previously experienced some speed-reading courses. Further, 67% of the subjects in each group indicated that they believed speed-reading courses can be effective. Thus, differential expectations of training effectiveness or past experience could not have contributed to the differences in reading rate changes that were obtained for the two populations.

An initial questionnaire assessed the subject's motivation to learn speed-reading techniques by asking how much time and money they would be willing to invest in a course. Volunteer and nonvolunteer subjects reported on the average that they would spend \$24 and \$29, respectively, and that they would

practice reading techniques 4.6 and 3.6 hours per week, respectively. Therefore, the two populations did not differ appreciably in their motivation to enhance their speed-reading skills.

In rating the experimenters' courtesy, competency, manners, and experimental conduct, all subjects tended to give favorable judgments. All group means were practically identical. On a 5-point scale (with 1 as the most favorable rating), the overall mean was 1.87. The present analogue to treatment is not complete. In clinical practice deprivation of treatment choice is rarely explicit. On the other hand, the analogue is more conservative, since agency-referred clients are often sent for treatment of a problem that they do not accept as requiring any personal change. In this study, both groups reported interest in working toward better reading skills. Thus the additional imposition of "the problem" in many clinical cases should yield stronger effects of the volunteer variable than was obtained here.

The overall findings of this experiment are consistent with the hypothesis that increased freedom of choice enhances performance in a change program. Within the limitations of the present study, this effect was more strongly demonstrated for the freedom to decide among training methods than for the difference in the source of referral to the experiment. Further tests of these hypotheses are needed to indicate whether generalization is warranted from the laboratory findings to long-term programs with clients who present a disturbing problem situation.

References

- Averill, J. R. Personal control over aversive stimuli and its relationship to stress. *Psychological Bulletin*, 1973, 80, 286-303.
- Brehm, J. W. *A theory of psychological reactance*. New York: Academic Press, 1966.
- Champlin, S. M., & Karoly, P. Role of contract negotiation in self-management of study time. *Psychological Reports*, 1975, 37, 724-726.
- Corah, N. L., & Boffa, J. Perceived control, self-observation, and response to aversive stimulation. *Journal of Personality and Social Psychology*, 1970, 16, 1-14.
- Cox, D. E., & Sippelle, C. N. Coercion in participation as a research subject. *American Psychologist*, 1971, 26, 726-728.
- Davison, G. C. Countercontrol in behavior modification. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology concepts and practice*. Champaign, Ill.: Research Press, 1973.
- Davison, G. C., & Valins, S. Maintenance of self-attributed and drug-attributed behavior change. *Journal of Personality and Social Psychology*, 1969, 11, 25-33.
- Gordon, R. M. Effects of volunteering and responsibility on the perceived value and effectiveness of a clinical treatment. *Journal of Consulting and Clinical Psychology*, 1976, 44, 799-801.
- Kanfer, F. H. Self-management methods. In F. H. Kanfer & A. P. Goldstein (Eds.), *Helping people change*. New York: Pergamon Press, 1975.
- Kanfer, F. H., & Seidner, M. Self-control: Factors enhancing tolerance of noxious stimulation. *Journal of Personality and Social Psychology*, 1973, 25, 381-389.
- Kelman, H. C. The rights of the subject in social research: An analysis in terms of relative power and legitimacy. *American Psychologist*, 1972, 27, 989-1016.
- Kopel, S., & Arkowitz, H. The role of attribution and self-perception in behavior change: Implications for behavior therapy. *Genetic Psychology Monographs*, 1975, 92, 175-212.
- Langer, R. J. The illusion of control. *Journal of Personality and Social Psychology*, 1975, 32, 311-328.
- Langer, E. J., Janis, I. L., & Wolfer, J. A. Reduction of psychological stress in surgical patients. *Journal of Experimental Social Psychology*, 1975, 11, 155-165.
- Langer, E., & Rodin, J. The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting. *Journal of Personality and Social Psychology*, 1976, 34, 191-198.
- Liem, G. R. Performance and satisfaction as affected by personal control over salient decisions. *Journal of Personality and Social Psychology*, 1975, 31, 232-240.
- Overall, J. E., & Woodward, J. A. Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, 1975, 82, 1, 85-86.
- Overall, J. E., & Woodward, J. A. Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin*, 1976, 83, 5, 776-777.
- Rosenthal, R., & Rosnow, R. L. *The volunteer subject*. New York: Wiley, 1975.
- Schultz, D. P. The human subject in psychological research. *Psychological Bulletin*, 1969, 72, 214-228.
- Stotland, E., & Blumenthal, A. The reduction of anxiety as a result of the expectation of making a choice. *Canadian Journal of Psychology*, 1964, 18, 139-145.
- Tucker, L. R., Damarin, F., & Messick, S. J. A basefree measure of change. *Psychometrika*, 1966, 31, 457-473.

Reduction of Autonomic Arousal in Alcoholics: A Comparison of Relaxation and Meditation Techniques

Jerry C. Parker and Gary S. Gilbert

Veterans Administration Hospital, Columbia, Missouri and
University of Missouri—Columbia

Richard W. Thoreson

University of Missouri—Columbia

To investigate and compare the effects of progressive relaxation training and meditation training on autonomic arousal in alcoholics, 30 subjects were selected from a population of alcoholics in a Veterans Administration hospital substance-abuse program. The subjects were randomly assigned to one of the following three experimental conditions: (a) progressive relaxation training group, (b) meditation training group, or (c) quiet rest control group. All groups met for 3 weeks during which state anxiety, blood pressure, heart rate, and spontaneous galvanic skin responses were measured. The measures were designed to assess the treatment effects following the first training session and at the end of the total training period. The results indicate that both progressive relaxation training and meditation training are useful for reducing blood pressure in alcoholics. In addition, significant differences between the groups in the effectiveness of the relaxation procedures were found. Meditation training induced blood pressure decreases at an earlier point in the 3-week training period and affected decreases in systolic blood pressure that progressive relaxation training did not. These results support the idea of considerable specificity of response to relaxation techniques.

Recent authors (Benson, Beary, & Carol, 1974; Bernstein & Borkovec, 1973) have discussed the usefulness of progressive relaxation training and meditation training as treatments for stress-related disorders. In general, these authors have portrayed the techniques as potentially effective treatments for a number of problems including hypertension, insomnia, tension headaches, and generalized anxiety.

Alcoholism and other forms of substance abuse are also disorders for which relaxation techniques may have utility. Preliminary support for this possibility has been provided by studies assessing the effects of meditation on self-reports of substance usage (Benson, 1974;

Benson & Wallace, 1972; Shaffi, Lavelly, & Jaffe, 1974, 1975; Winquist, 1973). Although the mechanism by which relaxation techniques might affect drinking behavior is unknown, two possibilities appear tenable. First, relaxation techniques might serve as a direct substitute for the reduced arousal that follows alcohol consumption. If this were the case, such techniques might be useful for alcoholics as coping techniques in stressful situations. Second, relaxation techniques might lead to a generalized state of lowered arousal. Such a state of lowered arousal would, in turn, tend to deprive alcohol of its hypothesized tension-reducing capacity, since a relatively lowered state of arousal would already exist. As a consequence, less direct reinforcement of drinking behavior would theoretically occur as a result of the use of alcohol.

Further credence in the possible role of relaxation techniques in the treatment of al-

This article is based on a doctoral dissertation submitted by the first author to the Graduate School of the University of Missouri—Columbia.

Requests for reprints should be sent to Jerry C. Parker, Psychology Service, Veterans Administration Hospital, Columbia, Missouri 65201.

cohol abuse is provided by studies investigating the relationship between anxiety, stress, and alcoholism. Steffen, Nathan, and Taylor (1974) found significant negative correlations between blood alcohol levels and electromyographic response in alcoholics, suggesting a tension-reducing effect of alcohol. Miller, Hersen, Eisler, and Hilsman (1974) compared alcoholics and social drinkers in a simulated interpersonal situation requiring assertive behavior. The alcoholics significantly increased their operant responses to obtain alcohol, but the social drinkers did not, suggesting that alcoholics may be more inclined to drink in response to social stress. Higgins and Marlatt (1975) studied the fear of interpersonal evaluation as a determinant of alcohol consumption in male social drinkers and found that subjects who expected to be evaluated in a second study drank significantly more alcohol than subjects who did not expect to be evaluated. The results of this study are suggestive of the possible role of stress, particularly when interpersonal in nature, in the development and maintenance of problematic drinking.

Although the tension-reduction hypothesis of alcoholism has not been firmly established (Cappell & Herman, 1972), the above studies are suggestive of possible dynamics involving anxiety, stress, and/or fear of interpersonal evaluation. Accordingly, relaxation techniques with reported success in reducing arousal (Benson et al., 1974; Bernstein & Borkovec, 1973) are worthy of consideration as a treatment strategy for alcoholics.

The purpose of the present study is to investigate the usefulness of relaxation techniques for reducing self-reported anxiety and autonomic arousal in alcoholics. Two divergent techniques, progressive relaxation training and meditation training, were included in the design for comparative purposes, since they may vary considerably in their effectiveness and in the dependent measures that they affect.

Method

Subjects

Subjects ($N = 30$) for this study were drawn from a population of male alcoholics on a substance abuse

unit of a Veterans Administration hospital. The mean age for the subjects was 45.1 years, and the mean educational level was 11.3 years. The subjects were generally of middle socioeconomic status, and the mean number of years of heavy drinking, as reported by the subjects, was 12.0. Subjects with evidence of psychosis, severe cerebral dysfunction, serious physical impairment, or illiteracy were omitted. In addition, subjects receiving psychotropic medications or having had previous relaxation or meditation training were not included. After the remaining patients were judged by the unit physician to be detoxified and after they had agreed to remain in the hospital for the duration of a 1-month treatment program, the State-Trait Anxiety Inventory (STAI), developed by Spielberger, Gorsuch, and Lushene (1970), was administered. Those patients with trait anxiety raw scores above 30 were selected for the study. The cutoff score was used because subjects with initially low anxiety would not permit an accurate assessment of the usefulness of the relaxation techniques. Two subjects were dropped on the basis of the cutoff score.

Groups

The selected patients were assigned by means of a table of random numbers to one of the following three treatment conditions:

1. Progressive relaxation training (PRT) group. The PRT treatment condition was operationalized by using the seven muscle group methodology of Bernstein and Borkovec (1973), which involves the systematic tensing and relaxing of various muscles. Subjects ($n = 10$) in this condition received 5 minutes of instruction in the rationale and exercises of PRT and then were exposed to 3 weeks of PRT practice, with the group meeting three times per week for $\frac{1}{2}$ hour each session. The first and last training sessions were individual sessions with the experimenter for the purpose of measurement. The middle sessions were conducted in a group setting with the 15-minute PRT instructions presented on a tape recorder and played in a quiet room with the lights turned low.

2. Meditation training (MT) group. The MT treatment condition was operationalized by using the methodology of Beary and Benson (1974). This methodology involves a comfortable position, a quiet environment, a passive attitude, and the silent repetition of the word *one* on every exhalation. Subjects ($n = 10$) in this condition received 5 minutes of instructions in the rationale and procedures of MT and then were exposed to 3 weeks of MT practice, with the group meeting three times per week for $\frac{1}{2}$ hour each session. The first and last training sessions were individual sessions with the experimenter for the purpose of measurement. The middle sessions were conducted in a group setting, with MT instructions presented on a tape recorder and played in a quiet room with the lights turned low. The first 2 minutes of the tape reviewed the MT tech-

nique and were followed by 13 minutes of silence, during which time the meditation was practiced.

3. Quiet rest control (QR) group. Subjects ($n = 10$) in this condition received 5 minutes of instruction in the desirability and benefits of relaxation, although no specific relaxation techniques were taught. Subjects were told simply to sit quietly with their eyes closed and to allow themselves to relax. They were then exposed to 3 weeks of QR sessions, with the group meeting three times per week for $\frac{1}{2}$ hour each session. The first and last QR sessions were individual sessions with the experimenter for the purpose of measurement. The middle sessions were conducted in a group setting, with the QR instructions presented on a tape recorder and played in a quiet room with the lights turned low. The first 2 minutes of the tape reviewed the QR instructions and were followed by 13 minutes of silence, during which time the quiet rest occurred.

Procedure

Subjects received a brief orientation interview by the experimenter, and permission for inclusion in a research study was obtained. The first session, an individual session with the experimenter, was conducted in a quiet room with the subject seated comfortably in a recliner facing away from the physiological recording equipment. The first part of the initial training session (5 minutes) was devoted to specific instruction in the particular treatment that the subject was to receive, and a 10-minute pretreatment measurement period immediately followed during which the physiological measures and the state anxiety data were collected. After this measurement period, the first 15-minute training tape was presented, and a 10-minute posttreatment measurement period followed in which the physiological measures and the state anxiety data were again collected. A brief period was provided after the second measurement period for questions from the subjects or further instructions, if needed.

The remaining training sessions, except for the final one, were conducted in a group setting over a 3-week treatment period. Each group meeting was $\frac{1}{2}$ hour in duration, with all groups meeting collectively for the first 15 minutes to control for motivational and expectancy variables. During this collective meeting, the rationale for relaxation therapy was briefly reviewed, and subjects were instructed to practice twice daily between group sessions.

The 15-minute tape-recorded training session followed, with each experimental group meeting separately in a quiet, dimly lighted room. The final training session, an individual session with the experimenter, again involved the collection of pretreatment and posttreatment measures.

Subjects who did not attend at least seven training sessions were dropped from the study. One subject from the QR group who had medical complications during his hospitalization was dropped for this reason. Another subject in the PRT group who dis-

charged himself from the hospital against medical advice was also dropped.

Measures

The following measures were designed to assess whether or not the treatments had immediate effects following the initial training session. To control for fluctuations due to the subjects' hospital activities, these measures were all collected in the mornings between the hours of 8:00 a.m. and 10:00 a.m.

1. The state anxiety portion of the STAI was administered during the pretreatment measurement period and the posttreatment measurement period of the initial training session.

2. Arterial blood pressure readings were taken on a standard medical manometer during the pretreatment measurement period and the posttreatment measurement period of the initial training session.

3. Heart rate measures were taken using the radial pulse method during the pretreatment measurement period and the posttreatment measurement period of the initial training session.

4. Skin conductance measures (spontaneous galvanic skin response; GSR) were recorded on a Lafayette Datagraph Model 76016 for a 2-minute interval during the pretreatment measurement period and the posttreatment measurement period of the initial training session. Spontaneous skin conductance fluctuations greater than .6 k Ω were scored by an independent observer unaware of the subjects' treatment condition. Beckman silver-silver chloride electrodes were attached to the first and third fingers of the left hand using Beckman electrode paste, and a 5-minute rest period was arranged during the pretreatment measurement period before the skin conductance measures were taken.

To assess the total training effects of the 3-week treatment program, STAI, blood pressure (BP), heart rate, and spontaneous GSR measures were also collected at the end of the final training session.

Results

The results of this study were analyzed in a two-way analysis of variance (Groups \times Trials) design with repeated measures on one factor. The pretreatment group means for state anxiety, BP, heart rate, and spontaneous GSR were not significantly different.

Session 1 Effects

State anxiety. The results from the 3×2 analysis of variance (ANOVA) on the STAI measures over Session 1 found the main effect for trials to be significant, $F(1, 27) = 43.49$, $p < .001$, indicating that the combined groups

Table 1
Mean State Anxiety Raw Scores Over the Initial Training Session and the 3-Week Training Period

Group	Measurement period		
	Session 1		Post final session
	Pre	Post	
Relaxation	45.5	35.2	28.8
Meditation	44.7	35.8	30.1
Control	46.5	41.6	36.6

reported less anxiety following Session 1, but no differences between the groups were found (see Table 1).

Blood pressure. The results from the 3×2 ANOVA on the systolic blood pressure measures over Session 1 found the interaction (Groups \times Trials) to be significant, $F(2, 27) = 4.68, p < .05$ (see Table 2).

For the systolic BP interaction, the group mean systolic BPs were significantly different from one another following their respective treatments, $F(2, 30) = 3.71, p < .05$. A Newman-Keuls probe found the MT group to be significantly lower than the PRT group ($r = 3, df = 30, p < .05$; see Winer, 1962, p. 80). The performance of the QR group, however, was not significantly different from that of

the PRT group or the MT group. The MT mean systolic BPs were significantly lower at the end of Session 1, $F(1, 27) = 11.52, p < .01$. The PRT and QR means were not significantly lower at the end of Session 1.

The results from the 3×2 ANOVA on the diastolic BP measures over Session 1 found the interaction (Groups \times Trials) to be significant, $F(2, 27) = 8.81, p < .01$ (see Table 2).

For the diastolic BP interaction, the MT diastolic BP mean was significantly lower at the end of Session 1, $F(1, 27) = 10.11, p < .01$. In contrast, the QR diastolic BP mean was significantly higher at the end of Session 1, $F(1, 27) = 6.77, p < .05$.

Heart rate. The results from the 3×2 ANOVA on the heart rate measures over Session 1 found the main effect for trials to be significant, $F(1, 27) = 26.70, p < .001$, indicating that the combined groups showed decreased heart rate following Session 1, but no differences between the groups were found (see Table 3).

GSR. The results from the 3×2 ANOVA on the spontaneous GSR measures over Session 1 found the main effect for trials to be significant, $F(1, 27) = 15.67, p < .001$, indicating that the combined groups showed decreased spontaneous GSRs following Session 1, but no differences between the groups were found (see Table 4).

Table 2
Mean Blood Pressures Over the Initial Training Session and the 3-Week Training Period

Group	Measurement period		
	Session 1		Post final session
	Pre	Post	
	Systolic		
Relaxation	118.0	117.5	115.0
Meditation	109.5	104.5	102.5
Control	109.5	110.5	119.0
	Diastolic		
Relaxation	84.0	81.5	78.0
Meditation	77.0	71.5	69.5
Control	71.5	76.0	80.5

Total Training Effects

State anxiety. The results from the 3×2 ANOVA on the STAI measures collected at the beginning of the first session and the end

Table 3
Mean Heart Rates/Min Over the Initial Training Session and the 3-Week Training Period

Group	Measurement period		
	Session 1		Post final session
	Pre	Post	
Relaxation	88.6	76.1	80.4
Meditation	86.8	80.6	76.9
Control	88.4	83.9	87.0

of the final session found the main effect for trials to be significant, $F(1, 27) = 50.80$, $p < .001$, indicating that the combined groups reported less anxiety at the end of the final training session than at the beginning, but no differences between the groups were found (see Table 1).

Blood pressure. The results from the 3×2 ANOVA on the systolic BP measures collected at the beginning of the first session and the end of the final session found both the main effect for groups, $F(2, 27) = 3.74$, $p < .05$, and the interaction (Groups \times Trials), $F(2, 27) = 11.16$, $p < .001$, to be significant (see Table 2).

For the systolic BP analysis, the group systolic BP means were significantly different from one another at the end of the final training session, $F(2, 38) = 7.54$, $p < .01$. A Newman-Keuls probe found that the MT group was significantly lower than both the PRT group ($r = 2$, $df = 38$, $p < .01$) and the QR group ($r = 3$, $df = 38$, $p < .01$). The MT mean systolic BP was significantly lower at the end of the final training session, $F(1, 27) = 7.38$, $p < .05$. The QR mean systolic BP was significantly higher at the end of the final training session, $F(1, 27) = 13.59$, $p < .001$.

The results from the 3×2 ANOVA on the diastolic BP measures collected at the beginning of the first session and the end of the final session found the interaction (Groups \times Trials) to be significant, $F(2, 27) = 12.90$, $p < .001$.

For the diastolic BP interaction, the PRT and MT mean diastolic BPs were significantly

lower at the end of the final training session, $F(1, 27) = 5.58$, $p < .05$, and, $F(1, 27) = 8.72$, $p < .01$, respectively. The QR mean diastolic BP was significantly higher at the end of the final training session, $F(1, 27) = 12.55$, $p < .01$.

Heart rate. The results from the 3×2 ANOVA on the heart rate measures collected at the beginning of the first session and the end of the final session found the main effect for trials to be significant, $F(1, 27) = 11.31$, $p < .01$, indicating that the combined groups had lower heart rates at the end of the final training session, but no differences between the groups were found (see Table 3).

GSR. The results from the 3×2 ANOVA on the spontaneous GSRs collected at the beginning of the first session and the end of the final session found no significant effects (see Table 4).

Discussion

Conclusions from Session 1 Analyses

With regard to whether PRT and MT have effects over a single treatment session, the results indicate that on several dependent variables, they were not significantly superior to control procedures. Although PRT and MT led to reports of less anxiety, slower heart rates, and fewer spontaneous GSRs, control procedures appeared to produce similar changes. These findings suggest that simply instructing subjects to "rest quietly" on their own is as effective as PRT or MT for inducing decreased self-reported anxiety, lowered heart rates, and decreased spontaneous GSRs.

The control procedures and PRT, however, were not as effective when systolic BP and diastolic BP were the dependent variables. The study of systolic BP over Session 1 indicates that the performance of the MT group decreased significantly during the first session, whereas, that of the PRT and QR groups remained approximately the same. Moreover, the analysis of diastolic BP changes over Session 1 indicates that the MT group decreased significantly during the first session, whereas the PRT group remained approximately the same, and the QR group increased significantly. In summary, PRT and

Table 4
Mean Spontaneous Galvanic Skin Responses
Over the Initial Training Session and the
3-Week Training Period

Group	Measurement period		
	Session 1		Post final session
	Pre	Post	
Relaxation	5.8	1.2	2.5
Meditation	7.1	2.7	3.0
Control	3.3	1.8	4.8

Note. Measures are the number of fluctuations greater than .6 kΩ per 2 minutes.

MT are not more effective than control procedures when self-reported anxiety, heart rates, and spontaneous GSRs are measured. However, MT is significantly more effective than PRT and control procedures when systolic BP and diastolic BP are measured.

Conclusions from Total Training Period Analyses

With regard to whether PRT and MT have effects over a 3-week training period, the results are generally analogous to the Session 1 analyses. All three groups showed decreases in self-reported anxiety and heart rates, but statistically significant differences between the groups were not found. The data suggest that subjects who are told to rest quietly relax as well as PRT and MT subjects when state anxiety scores and heart rates are the dependent measures. No significant training effect was demonstrated for the spontaneous GSR measure, indicating that the Session 1 effects did not last over the entire training period.

Group differences were again found on the systolic BP and the diastolic BP measures. The systolic BP analysis showed that the MT group decreased significantly, the PRT group remained approximately the same, and the QR group *increased* significantly. The diastolic BP analysis showed that both the PRT and the MT groups decreased significantly while the QR group *increased* significantly. In summary, PRT and MT do not affect self-reported anxiety or heart rates over a 3-week training period, but they do affect systolic and diastolic BP in the direction of decreased arousal.

Implications

The results of this study suggest that persons have considerable "control" over the relaxation process, even when formal relaxation techniques are not being used. In terms of heart rates and spontaneous GSRs, the QR condition affected significant decreases on several measures. This finding is in contrast to the findings of several other studies (Beary & Benson, 1974; Wallace, 1970; Wallace & Benson, 1972) and may stem from the fact that expectancy, motivation, and attention ef-

fects were carefully controlled in the present research. Since the groups in this study met collectively when the rationale for relaxation was explained, the QR subjects had the same "expectancy" for benefits as did the experimental groups. Accordingly, the QR subjects were provided with a set that elicited a strong "desire" to relax.

A perusal of the typical research in the area of relaxation shows that such careful control of expectancy and attention is not usually the case. For example, in the typical ABA designs, trained meditators are instructed to "sit with eyes closed," then to "meditate," and then to "sit with eyes closed" again. The independent variable of meditation is manipulated in such designs, but expectancy is too. Subjects are quite probably given subtle cues indicating that they are supposed to "try" in the meditation condition, at least more so than when "sitting with eyes closed." In the Wallace and Benson (1972) research, the subjects were fully aware of the purpose of the study and had an investment in proving that meditation was effective. The present study suggests that expectancy and motivation are important variables that must be controlled in relaxation research, since subjects can affect considerable decreases in arousal on certain dependent variables with only positive expectancies and motivations operating.

A second implication of this study is that a single dependent variable is not likely to be sufficient for satisfactorily evaluating the effects of a relaxation procedure. Both systolic and diastolic BP differentiated between the groups in this study, but heart rates, spontaneous GSRs, and self-reports did not. Several studies in the literature of both relaxation training and meditation have used self-reports exclusively, and such reliance on a single dependent measure can sometimes be misleading. In the present study, systolic and diastolic BP measures were the most sensitive to differences between the groups.

A third implication is that PRT and MT differ substantially in their efficiency for inducing decreased autonomic arousal as measured by BP. In the systolic and diastolic BP analyses over Session 1, the MT group showed significant decreases, whereas, the PRT group

did not. In the systolic and diastolic BP analyses across the total training period, the MT group continued to show significant decreases, whereas the PRT group showed a significant decrease in diastolic BP only. On the basis of these findings, MT is a more efficient relaxation technique than is PRT when BP is a relevant variable.

A fourth, most intriguing, implication of this research is that efforts at relaxation can sometimes be stressful. In the analysis of diastolic BP over Session 1, the QR group showed a significant *increase* in arousal. Although heart rate, spontaneous GSR, and state anxiety scores decreased in association with the QR procedure, diastolic BP increased. This finding leads to the hypothesis that "cognitive activity" may increase in a stimulus-deprived situation as was found in the QR condition. Whereas the experimental groups had an activity to perform (either mental or physical exercises), the control group did not. All three groups appear to have received the benefits of a sedentary physical state, but only the experimental groups received the benefits associated with restricted attention. The restricted attention apparently reduced the likelihood of experimental subjects thinking about current problems and/or anxiety-arousing situations as seemed to occur with the QR control subjects. The experimental groups may have escaped such cognitive activity by the demands of the relaxation task. The findings of this study suggest that the MT procedure was the most effective experimental condition in a number of the analyses, and the concept of restricted attention may help explain these results.

Finally, with regard to the question of whether relaxation techniques are potentially useful in the treatment of alcoholism, the data are encouraging. The alcoholic subjects in this study clearly experienced decreases in autonomic arousal and self-reported anxiety, with MT tending to be the most efficient treatment condition. Using only a 5-minute instruction period, significant decreases in arousal were obtained during the first training session in both the PRT and MT groups. In general, these decreases were maintained over a 3-week training period with only group support. Such

benefits are at a small cost in terms of staff time, which makes the relaxation techniques practical for substance abuse programs. Also, the alcoholics in this study accepted the relaxation procedures with interest and cooperation. None of the subjects refused to participate or expressed dissatisfaction with the relaxation procedures.

As a cautionary note, demonstrating the efficiency of relaxation techniques for reducing autonomic arousal in alcoholics is not equivalent to demonstrating their usefulness as direct treatments for alcoholism. Long-term studies are required to assess whether alcoholics can be trained to use relaxation techniques on a regular basis over an extended period of time. Moreover, will such extended usage result in a generalized reduction in autonomic arousal and/or decreased rates of alcohol consumption?

References

- Beary, J., & Benson, H. A simple psychophysiologic technique which elicits the hypometabolic changes of the relaxation responses. *Psychosomatic Medicine*, 1974, 36, 115-120.
- Benson, H. Decreased alcohol intake associated with the practice of meditation: A retrospective investigation. *Annals of the New York Academy of Sciences*, 1974, 233, 174-177.
- Benson, H., Beary, J., & Carol, M. The relaxation response. *Psychiatry*, 1974, 37, 37-46.
- Benson, H., & Wallace, R. Decreased drug abuse with transcendental meditation—A study of 1,862 subjects. In C. Aarafonotis (Ed.), *Drug abuse: Proceedings of the international conference*. Philadelphia: Lea & Febiger, 1972.
- Bernstein, D., & Borkovec, T. *Progressive relaxation training: A manual for the helping professions*. Champaign, Ill.: Research Press, 1973.
- Cappell, H., & Herman, C. Alcohol and tension reduction: A review. *Quarterly Journal of Studies on Alcohol*, 1972, 33, 33-64.
- Higgins, R., & Marlatt, G. Fear of interpersonal evaluation as a determinant of alcohol consumption in male social drinkers. *Journal of Abnormal Psychology*, 1975, 84, 644-651.
- Miller, P., Hersen, M., Eisler, R., & Hilsman, G. Effects of social stress on operant drinking of alcoholics and social drinkers. *Behaviour Research and Therapy*, 1974, 12, 67-72.
- Shafii, M., Lavelly, R., & Jaffe, R. Meditation and marijuana. *American Journal of Psychiatry*, 1974, 131, 60-63.
- Shafii, M., Lavelly, R., & Jaffe, R. Meditation and the prevention of alcohol abuse. *American Journal of Psychiatry*, 1975, 132, 942-945.

- Spielberger, C., Gorsuch, R., & Lushene, R. *Manual for the State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Steffen, J., Nathan, P., & Taylor, H. Tension-reducing effects of alcohol: Further evidence and some methodological considerations. *Journal of Abnormal Psychology*, 1974, 83, 542-547.
- Wallace, R. Physiological effects of transcendental meditation. *Science*, 1970, 167, 1751-1754.
- Wallace, R., & Benson, H. The physiology of meditation. *Scientific American*, 1972, 226, 84-90.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.
- Winquist, W. The effect of the regular practice of transcendental meditation on students involved in the regular use of hallucinogenic and "hard" drugs. In D. Kanellakos & P. Ferguson (Eds.), *The psychobiology of transcendental meditation: An annotated bibliography*. Los Angeles: Maharishi International University Press, 1973.

Received April 27, 1977 ■

Sex and Worker Acceptance of a Former Mental Patient

Amerigo Farina, Pauline J. Murray, and Thomas Groh
University of Connecticut

This is the fifth in a series of studies measuring the acceptance accorded to former mental patients. The procedure was to send a confederate in the guise of a job applicant to be evaluated by a worker already on the job. The workers were told either that the applicant was an ex-mental patient or that he was an ordinary applicant, and for each condition the confederate was calm for half of the subjects and nervous for the rest. The studies revealed that women are more accepting of former patients than men and that men are more accepting of female than male ex-patients. Nervous applicants were rejected by workers of both sexes.

Persons who experience difficulty in adjustment are viewed in a grossly unfavorable way by virtually all members of our society (Nunnally, 1961). Opinions are especially negative toward those who are hospitalized, apparently because hospitalization indicates very serious mental problems (Lawner, 1966). This leads to the gloomy conclusion that those who are least able to deal with thorny interpersonal situations and are hospitalized for this reason will find their problems magnified when they return to the community. Research into various facets of interpersonal interaction shows that this is precisely what happens (Farina, Gliha, Boudreau, Allen, & Sherman, 1971; Farina, Holland, & Ring, 1966; Farina, Thaw, Lovern, & Mangone, 1974).

A particularly important area for community adjustment is employment, since self-esteem and social status are closely tied to the job one has. Moreover, having enough money for even basic necessities such as food will typically entail working, and if an ex-patient cannot find a job, the patient may be forced to return to the hospital. The research that has been done on job finding has been

limited to male subjects. However, that research makes it amply clear that males are denied jobs if they have been in a mental hospital. Employers will openly *say* that they do not like hiring ex-mental patients (Olshansky, Grob, & Malamud, 1958). With actual behaviors, it was found that employment interviewers estimated a male applicant's probability of getting a job as significantly lower, and they were reliably less friendly toward him when he revealed a history of mental illness in comparison to when he did not (Farina & Felner, 1973).

In view of these data, some wholly unexpected findings were obtained in a series of four studies, all using the same procedure (Farina, Felner, & Boudreau, 1973; Farina & Hagelauer, 1975). Female department store clerks met a female confederate in the guise of a job applicant and were asked to evaluate her as a potential co-worker. Half were told that the applicant was an ex-mental patient, and the rest were told that she was an ordinary job seeker. Unlike all prior studies of males, the women were no less accepting of the ex-mental patient than of the normal applicant. The study was replicated with female workers in a hospital, using a different female confederate, and the identical results were obtained. There followed another replication at the same hospital, but this time the subjects were male workers who met a male confederate. It was then found that males

The present study was, in part, financed by Research Grant GS-31030X from the National Science Foundation.

Requests for reprints should be sent to Amerigo Farina, Department of Psychology, U-30, University of Connecticut, Storrs, Connecticut 06268.

strongly rejected the applicant with a history of mental illness. To determine if it was sex of the subject or sex of the mental patient that accounted for the differences, a fourth replication was carried out. A new group of female department store clerks met a different male confederate, and no difference was found in degree of acceptance accorded to the applicant whether he was perceived as an ex-patient or as normal.

Attitude studies hardly explain these clear sex differences in acceptance of former mental patients. Nunnally (1961) found no sex differences in attitudes in any of the numerous studies that he has done. In a later attitude study (Farina et al., 1973), it was found that males in comparison to females expressed a greater liking for and a greater willingness to work with ex-mental patients. This kind of incongruity between attitudes and behaviors is not peculiar to the topic of mental illness. It has been found in a number of areas (e.g., Farina & Holzberg, 1967; Kutner, Wilkins, & Yarrow, 1952).

Thus, prior behavioral studies have shown that females are perfectly willing to accept either a male or a female as a co-worker even if the applicant has been mentally ill. Male workers, however, strongly reject a male applicant if he has a history of mental illness. How males would act toward a female applicant who has been in a mental hospital has not been determined, and this is one purpose of the present study. In the prior studies an additional variable, being tense and nervous or being calm, was also investigated. With half of the subjects in the mentally ill condition and half in the control condition, the "applicant" acted visibly tense and nervous. Calm and relaxed behavior was displayed with the remaining subjects. In all four studies the nervous applicant was unambiguously rejected. This manipulation is also carried out in this investigation to determine if tense people are disliked generally or whether this effect, too, is sex related.

Method

The subjects of the study were 48 males employed in the physical plant division of a state university. The nature of the work they did was quite variable, ranging from supervision of heating plant operations

to janitorial work. They were contacted by a supervisor and were informed that the possibility of using workers to evaluate job applicants was being investigated. Therefore, they were told that workers were needed to interview and give their opinion of a potential co-worker, and they were asked to volunteer their help during regular working hours. Those who volunteered were contacted, and a time was set for the interview to take place. All but 1 man who volunteered actually became subjects of the study, and 12 were randomly assigned to each of the four cells. The single exception was a man who expressed the conviction that we were trying to fire people like him, and he was excused from the study.

The workers were seen in offices belonging to the physical plant division. It was explained to them that workers already on the job knew that job best and, therefore, they would meet a job applicant and evaluate how well the new person would do if placed in the same department in which they worked. They were further told that the university wanted to see how disadvantaged people (former mental patients) would do on certain jobs. To find out, they were told that some workers would meet former mental patients who were interested in working for the university and that other workers would meet ordinary applicants to provide comparison data by means of which the suitability of ex-patients could be judged. To avert suspicion in the event that the employees compared notes, they were told that one applicant might be seen by several workers.

Each subject was asked to talk to the applicant and to form a basic impression about her. This was stressed as most important, since later, in private, he was to indicate how he judged the applicant would do if hired. He was also asked to describe the essentials of his work to help her decide about accepting the job, if it were offered. As part of a brief background statement about the applicant, half of the workers were then told that the person they were about to meet was a former mental patient, whereas the rest were informed that she was an ordinary job applicant. Actually, everybody met the same confederate,¹ a female undergraduate student in her early 20s. She was introduced with different names, and she changed clothing several times in a given day to avert suspicion. The workers generally seemed pleased to participate in the project and appeared to believe the experimenter.

The confederate related the same personal history in all conditions. She reported having graduated from high school and working as a clerk in a hardware store. She also said that she had worked in a grocery store and had been employed as a secretary. However, with half of the workers in the mental patient condition and half in the control condition, she behaved in a calm, relaxed manner, whereas with the rest she was tense and anxious. The be-

¹ We would like to thank Maureen Kohler for doing this work.

havior selected to indicate anxiety was the same as was used in the prior four studies. In the anxious condition the confederate seldom looked the employee in the eye, only occasionally stealing a glance. She kept her head down while frequently wringing her hands, and periodically she swallowed as if her throat were dry. To avoid possible changes in behavior as a function of believing the worker considered her normal or blemished by a history of mental illness, she was kept ignorant of what the subjects had been told.

After the confederate had described her work experience, the employee was instructed to describe his job. The applicant was then thanked and dismissed, and a postexperimental questionnaire was read to the worker, who responded to the items verbally. None of the workers seemed suspicious about any aspect of the study. We believe that participation in the study during regular working hours and at the request of the supervisor helped to avert skepticism. We further believe that the effort being made by the university during that period to hire from populations underrepresented by the employees also helped. The true nature of the study was never revealed both to insure subject naiveté and because disclosure seemed more harmful than nondisclosure.

Results and Discussion

A central purpose of the present study was to determine the reception that would be given to a female former mental patient by male workers. To do this, the 15 items in the postexperimental questionnaire were analyzed using two-factor analyses of variance, with one factor being history of the applicant (mental illness—no mental illness) and the other component being the confederate behavior (tense—calm). For one of these items, a significant main effect ($p < .05$) for mental illness was found. When asked to describe the applicant's assets, the workers perceived her as having fewer assets if she had been in a mental hospital than if she had not. Although this finding was close to chance level (1 significant item in 15), it was anticipated that not all items would be equally revealing about unfavorable dispositions toward ex-patients. In our society people are expected to be helpful toward unfortunate others, like mental patients, rather than doing them harm. Consequently, as in the prior studies, the workers were not expected to express disfavor openly, such as by saying they would not get along with the applicant in the mental patient condition. However, the expression

of negative feelings was expected on items on which it was not obvious that the respondent was registering personal disfavor. One item viewed as especially subtle was the one inquiring about the assets possessed by the applicant. This reasoning is described in detail, and empirical support for it is presented in another publication (Farina, Chapnick, Chapnick, & Misiti, 1972). Therefore, even though only one item differed between conditions, the fact that it was a subtle item suggests that the history of hospitalization did lead the workers to reject the female applicant. However, this rejection is much milder than that shown toward a male applicant in a previous study (Farina et al., 1973).

There are also some significant interactions which indicate that these subjects have an unfavorable view of a female former patient. In the calm condition the confederate was described as significantly more tense when she was presented as a patient than when she was described as an ordinary job applicant. On the other hand, when nervous and tense, it was the normal applicant who was rated as more tense ($p < .05$). It seems that the normal applicant is expected to be calm, and when she is tense her nervousness is very salient. In contrast, the ex-mental patient is expected to be nervous, and she is seen as tense even when objectively the tension is not there. A similar pattern was found for the item asking how well the applicant would do the job if hired. A reliable interaction ($p < .05$) indicates that the normal person is expected to do better when calm, whereas a former patient is expected to do better when nervous. Possibly, better performance is anticipated and people are liked more when they behave in accordance with stereotypic beliefs. One of the studies reported by Farina et al. (1973, p. 366) strengthens this possibility, since the same results were found.

The findings concerning the calm—tense variable were consistent and quite clear. When tense, the confederate is expected to get along less well with other workers ($p < .05$), she is perceived as having fewer assets ($p < .001$) and more liabilities ($p < .001$), she is judged to be less reliable ($p < .05$), and she is thought to be less well adjusted

($p < .001$) and valuable ($p < .05$) than when she is calm.

Before considering the meaning of these results, we need to take note of the methodology used in this and the similar prior studies, since this bears on the conclusion we can draw. In each study, the subjects were a reasonably random sample of workers in that setting, and so we believe that other workers drawn from comparable populations would behave like our subjects if they were to face the same conditions. However, only one confederate was used in each study. The major problem with such a procedure was clearly recognized by Brunswik (1947) and forcefully described by Hammond (1948). In brief, if we want to determine how workers respond to an ex-mental patient, we need both a representative sample of workers and a representative sample of ex-mental patients. Of course, we can conclude that workers react unfavorably to at least one person—the confederate we used. But another individual, whose physique and personality are different, might not elicit the same reaction. Data pertinent to this issue were reported by Farina, Thaw, Felner, and Hust (1976). Those researchers examined how the social impact of stigmatizing conditions (including mental illness) was influenced by individual differences among people. Four confederates played the role of a normal or stigmatized person with subjects drawn from the same population. A given subject saw only one confederate in only one of the three roles. It was found that the effect of the stigmatizing conditions was reliably influenced by the individual characteristics of each confederate.

On the other hand, Farina et al. (1976) also found response patterns that were the same for all confederates; for example, when perceived as mentally retarded, each confederate was given reliably less painful shocks than when perceived as normal. Thus, individual differences among people do not necessarily invalidate all findings obtained with just one person. More important, the present study is the fifth in a series, all using identical procedures. The five experiments contained five different confederates, two males and three females, and were done in three different places of employment. Yet, the findings

are extremely consistent. For the calm-tense variable, we find that the confederate, when nervous, is always evaluated much less favorably. The mentally ill-normal results are also consistent, with males evaluating the confederate in the ex-mental patient condition less favorably than in the control condition and women responding the same in the two conditions.

What conclusions can we reasonably draw from this and the earlier studies? There appear to be two rather clear sex differences in the way that mental patients are treated. First, men appear to be unfavorably disposed toward members of either sex if they have been mentally ill, whereas women fully accept such a person. And second, focusing on the victim of the mental illness, a male is treated more poorly than a female even when they have identical psychiatric histories. The results from this series of five studies are surprising in view of the past studies that uniformly report that there are no sex differences of the sort that we have found. Leaving inconsistencies between the present and past studies aside for the moment, how are we to understand our findings? There are reports in the literature that suggest explanations.

Parsons and Bales (1955) have asserted that women are more concerned with ongoing interpersonal relationships than are men, whereas men focus more on goals in the future. The confederate acted in the same way both as a normal and as an ex-mental patient; women, perhaps, respond similarly in the two conditions because they are more influenced by the immediate behavior than by the cues about future threats from the former patient. On the other hand, men are more concerned about the future and may have rejected the applicant with the psychiatric history because of fears such as of a disruption in their work careers. As for the preference men show for female in comparison to male ex-mental patients, a possible explanation is that males are thought more likely to be aggressive and disruptive once hired. It has been reported that males are viewed as more disposed than females to react to stress with aggressive behavior

(Coie, Pennington, & Buckley, 1974). However, the evidence in support of these explanations is of uncertain value, since it is based on surveys of opinions. It appears likely that it is because attitudinal measures were used in the prior studies of social reactions to mental patients that the sex differences we found were so unexpected. By now it should be very clear that the attitudes that people express and what they actually do are not necessarily consistent. Hence, if we are interested in behavior, we should measure behavior as directly as possible.

The finding that nervous people are so unequivocally and strongly rejected has both theoretical and practical implications. On the theoretical level, it seems important to know why tense people are responded to so unfavorably. Perhaps such a response is limited to circumstances in which someone is to be judged as a worker and means only that anxious people are expected to do a poor job. But it seems likely that this reaction is much more general, and it may mean that nervous people evoke memories of negative events or promise to bring trouble in the future. Whatever is responsible, we can expect that tense people will be less readily accepted than calm individuals, and we can improve their social relationships if we can reduce the visibility of their tension. If mental patients are especially tense, as is widely believed, it might be particularly important to their readjustment at home if they can be helped to look more calm. Conceivably, the nervousness displayed by ex-patients is in part responsible for the difficulties that they encounter in finding a job.

References

- Brunswick, E. *Systematic and representative design of psychological experiments*. Berkeley: University of California Press, 1947.
- Coie, J. D., Pennington, B. F., & Buckley, H. H. Effects of situational stress and sex roles on the attribution of psychological disorder. *Journal of Consulting and Clinical Psychology*, 1974, 42, 559-568.
- Farina, A., Chapnick, B., Chapnick, J., & Misiti, R. Political views and interpersonal behavior. *Journal of Personality and Social Psychology*, 1972, 22, 273-278.
- Farina, A., & Felner, R. D. Employment interviewer reactions to former mental patients. *Journal of Abnormal Psychology*, 1973, 82, 268-272.
- Farina, A., Felner, R. D., & Boudreau, L. A. Reactions of workers to male and female mental patient job applicants. *Journal of Consulting and Clinical Psychology*, 1973, 41, 363-372.
- Farina, A., Gliha, D., Boudreau, L. A., Allen, J. B., & Sherman, M. Mental illness and the impact of believing others know about it. *Journal of Abnormal Psychology*, 1971, 77, 1-5.
- Farina, A., & Hagelauer, H. D. Sex and mental illness: The generosity of females. *Journal of Consulting and Clinical Psychology*, 1975, 43, 122.
- Farina, A., Holland, C. H., & Ring, K. The role of stigma and set in interpersonal interaction. *Journal of Abnormal Psychology*, 1966, 71, 421-428.
- Farina, A., & Holzberg, J. D. Attitudes and behaviors of fathers and mothers of male schizophrenic patients. *Journal of Abnormal Psychology*, 1967, 72, 381-387.
- Farina, A., Thaw, J., Felner, R. D., & Hust, B. E. Some interpersonal consequences of being mentally ill or mentally retarded. *American Journal of Mental Deficiency*, 1976, 80, 414-422.
- Farina, A., Thaw, J., Lovern, J. D., & Mangone, D. People's reactions to a former patient moving to their neighborhood. *Journal of Community Psychology*, 1974, 2, 108-112.
- Hammond, K. R. Subject and object sampling—A note. *Psychological Bulletin*, 1948, 45, 530-533.
- Kutner, B., Wilkins, C., & Yarrow, P. R. Verbal attitudes and overt behavior involving racial prejudice. *Journal of Abnormal and Social Psychology*, 1952, 47, 649-652.
- Lawner, P. *Unfavorable attitudes and behavior toward those bearing mental illness stigmas*. Unpublished master's thesis, University of Connecticut, 1966.
- Nunnally, J. C., Jr. *Popular conceptions of mental health*. New York: Holt, Rinehart & Winston, 1961.
- Olschansky, S., Grob, S., & Malamud, I. T. Employer's attitudes and practices in the hiring of ex-mental patients. *Mental Hygiene*, 1958, 42, 391-401.
- Parsons, T., & Bales, R. F. *Family socialization and interaction process*. Glencoe, Ill.: Free Press, 1955.

Received May 2, 1977 ■

Prospects for Faking Believable Deficits on Neuropsychological Testing

Robert K. Heaton, Harold H. Smith, Jr., Ralph A. W. Lehman,
and Arthur T. Vogt
University of Colorado School of Medicine

This study compared the results of 16 volunteer malingers with those of 16 cooperative, nonlitigating head-trauma patients on the Wechsler Adult Intelligence Scale, the Halstead-Reitan battery, and the Minnesota Multiphasic Personality Inventory (MMPI). The overall level of ability impairment shown by the malingers equaled that of the head-injury group, but different patterns of strengths and deficits were produced by the two groups on testing. The malingers also showed more severe personality disturbance on the MMPI. The test protocols were sent to 10 neuropsychologists, who made "blind" judgments as to whether each was probably produced by a malingerer or by a real head-injury patient. Neuropsychologists' diagnostic accuracies ranged from chance-level prediction to about 20% better than chance. Discriminant functions based on the neuropsychological test results and the MMPI, respectively, correctly classified 100% and 94% of subjects in both groups. In another large sample of head-injury patients, those who were involved in court actions and/or gave clinical evidence of faking were more likely to be classified as malingers by the discriminant functions.

Neuropsychological tests are widely used in clinical settings to help diagnose brain lesions. Their role in this context is usually ancillary. That is, they help predict the presence and nature of neurologic conditions, which must then be confirmed by procedures involving more cost, risk, and/or discomfort. On the other hand, because they directly measure abilities affected by brain damage, neuropsychological tests have a more definitive role when the questions being asked deal specifically with the behavioral consequences of neurologic conditions. Such consequences can be important in themselves, not merely as symptoms contributing to a medical diagnosis. For

the many brain-damaged patients whose conditions are neither grossly incapacitating nor immediately life threatening, it is important to know which abilities have been affected, the severity of the various deficits, and their probable impacts on the patients' everyday functioning (e.g., on social relationships and potential for working and living independently). In some civil and criminal court proceedings, there is also a need to define as precisely as possible the effects of brain lesions on adaptive abilities. In such cases a dollar value is to be placed on the disability resulting from a head injury, or legal competency is to be decided.

This project was supported in part by Biomedical Research Support Grant RR-05357 awarded by the Biomedical Research Support Grant Program, Division of Research Resources, National Institutes of Health.

Harold H. Smith, Jr., is currently in independent practice in Seminole, Florida.

Requests for reprints should be sent to R. K. Heaton, Mail Container C256, University of Colorado Medical Center, 4200 East Ninth Avenue, Denver, Colorado 80262.

To be valid for any of the purposes mentioned above, neuropsychological testing requires adequate effort on the part of the patient. Most people are inclined to do their best on tests, and most patients have more to gain by appearing capable than by emphasizing deficits. This is not always true, however, particularly when the test results are to be used to justify compensation or other claims. Thus, in testifying before the courts as an expert witness, the clinical neuropsychologist

chologist is frequently asked a difficult question: "Is it possible to exaggerate deficits on these tests or even to fake deficits that do not exist?" In some cases one can argue convincingly that the particular pattern of test results makes good sense neurologically, and that it is therefore unlikely that the patient was malingering. However, few (if any) clinicians have had experience with known malingerers, and no research has shown whether malingerers' neuropsychological test scores can be distinguished from those of nonmalingerers' brain-damaged patients.

The Minnesota Multiphasic Personality Inventory (MMPI) is often used in conjunction with neuropsychological tests to help determine the types and degrees of emotional disturbances that may be associated with neurologic impairment (Boll, Heaton, & Reitan, 1974; Dikmen & Reitan, 1977). MMPI scales and formulas have been developed for detecting conscious malingering and other test-taking attitudes that may affect MMPI profile validity or serve as moderator variables in profile interpretation (Anthony, 1971; Cofer, Chance, & Judson, 1949; Gough, 1947, 1950, 1954; Osborne, 1970). It is not known whether patients tend to display the same test-taking attitudes on the MMPI and on neuropsychological tests; if they do, MMPI signs of malingering might alert clinicians to the possible invalidity of neuropsychological protocols produced by the same patients. Also, there is some evidence that the MMPI may be helpful in identifying patients with functionally based neurological complaints (Shaw & Matthews, 1965).

The present study compares the results of some volunteer malingerers with those of nonlitigating head-injured patients on the MMPI and a detailed battery of neuropsychological tests. In addition, these test results were sent to 10 neuropsychologists to determine whether they could judge which protocols were produced by malingerers and which were produced by the patients with real head injuries.

Method

Subjects

The head-injury group was composed of 13 males and 3 females, with a mean age of 26.7 ($SD = 6.5$)

years and a mean of 11.9 ($SD = 1.6$) years of education. All of these patients had been referred for clinical neuropsychological evaluations, and all were considered to have put forth adequate effort on the testing. All had documented histories of traumatic head injuries, followed by at least 12 hours of unconsciousness, and had no history of any other neurological illness. At the times of testing, all of the patients had residual neurological deficits, but none had peripheral injuries to their upper extremities that would interfere with test performance. Also, none of these individuals were involved in civil or criminal court actions, and none were applying for disability income support.

The malingering subjects were recruited by the second author from his neighborhood, college classes, and church. These subjects were told that they would be given \$25 for taking the neuropsychological test battery and a \$5 bonus if they were successful fakers. Actually, all subjects were paid the \$30 regardless of how much they faked. Twenty subjects agreed to participate in the study as malingerers, but only 16 gave any evidence of actually malingering on the tests. The other 4 individuals earned normal scores on the entire neuropsychological test battery and were eliminated from the study.¹ There are two reasons for omitting them. First, the normal test scores would not have justified compensation in a real court case. Second, the purpose of this study was to learn whether patterns of neuropsychological deficits due to head injury can be distinguished from those faked by malingerers. The distinction between real deficits and no deficits is of little interest in this context.

The malingerers who were retained in the study included 11 males and 5 females. Their mean age of 24.4 years ($SD = 7.5$) was not significantly different from that of the real head-injury group, $t(30) = .94$, $p > .05$. Also, their mean of 12.9 years ($SD = 2.4$) of education was comparable to that of the head-injury group, $t(30) = 1.39$, $p > .05$. To provide an estimate of their actual intelligence, prior to their testing as malingerers, these subjects were administered the Shipley-Hartford Intelligence Scale (Shipley, 1940) under standard conditions; that is, they were asked to do their best. Using the system proposed by Paulson and Lin (1970) for predicting Wechsler Adult Intelligence Scale (WAIS) Full Scale IQ values from Shipley scores, we obtained a group mean of 113.2 ($SD = 7.5$).

Tests

Neuropsychological evaluations were administered by experienced technicians who were trained and

¹ One of these four individuals had started to fake, but he was so obvious about it that the technician severely reprimanded him and threatened to call the attorney who had supposedly referred him for testing. From that point on, the subject earned above average scores on all tests given.

supervised by the first author. The evaluations consisted of all 11 subtests of the WAIS (Wechsler, 1955) and all of the tests normally included in the Halstead-Reitan Neuropsychological Test Battery for adults. These latter tests, which are described in detail in Reitan and Davison (1974), include the Category Test; the Tactual Performance Test; the Speech-Sounds Perception Test; the Seashore Rhythm Test; the Finger Oscillation Test; Parts A and B of the Trial Making Test; the Reitan-Kl ve Sensory-Perceptual Examination, with the Tactile Form Recognition Test replacing Tactile Coin Recognition as the measure of dysstereognosis; the Aphasia Screening Test, including the testing of visuconstructive abilities; measures of grip strength, obtained with a hand dynamometer; the Grooved Pegboard and Static Steadiness Tests from the Kl ve-Matthews Motor Steadiness Battery; and the Reitan-Kl ve Lateral Dominance Examination. The MMPI was also administered to provide objective personality measures that might be differentially associated with real brain damage versus deliberate attempts to exaggerate pathology.

Numerous published studies have shown that the neuropsychological test battery used in this study is sensitive to focal and diffuse cerebral lesions caused by diverse neurologic conditions. (See reviews of this literature in Kl ve, 1974, and Russell, Neuringer, & Goldstein, 1970.) In addition, this battery is well suited to the task of cataloging residual strengths and deficits after a traumatic head injury, because it samples broadly the adaptive abilities that can be affected by cerebral lesions, such as, sensory, motor, cognitive, language, visuospatial, as well as other mental abilities. An additional advantage of the test battery approach is that it permits analysis of both the level and the pattern of test performance in making diagnostic inferences (Reitan, 1966). Given the test scores from the Halstead-Reitan battery, experienced clinicians have been able to use these complementary diagnostic methods to infer not only the presence and location of cerebral lesions but also their etiologies (Filskov & Goldstein, 1974; Reitan, 1964). Even though it may be very difficult for a clinician to discern whether one or a few poor test scores have been faked, a consideration of the pattern of strengths and deficits shown on more comprehensive testing may make such discrimination possible; that is, to fool the clinician on the test battery, the malingerer must show a pattern of results that is similar to patterns of scores earned by head-injury patients with real deficits.

Judges

Ten neuropsychologists provided independent "blind" ratings as to whether they thought each test protocol was produced by a malingerer or by a nonmalingerer head-injury patient.² The second and fourth authors participated as judges. However, at that time they had no knowledge of the numbers of subjects in each group and no previous

exposure to the test results of any of the subjects. The 10 judges differed greatly with respect to previous experience in interpreting the Halstead-Reitan battery: The range was 8 weeks to 18 years full-time equivalent, with 5 of the judges having 4 or more years of experience. Judges with varying experience were sought in order to assess the possible role of experience in increasing diagnostic accuracy.

Procedure

The technicians who tested the malingering subjects were not told about this study until after the data collection phase was completed. To replicate the usual testing situation as closely as possible, it was considered necessary for the technicians to test these subjects as real clinical patients. Therefore, these subjects were scheduled as litigating patients, referred by one of two Denver attorneys who had previously sent real patients for testing.³

In preparation for their neuropsychological evaluations, the volunteer malingerers were asked to pretend that they had suffered head injuries in accidents caused by other persons. Subjects were to consider themselves involved in litigation to determine how much financial compensation they would obtain from the persons responsible for the accidents or from the insurance companies involved. They were told to imagine that their everyday functioning (e.g., in school and vocational activities) had been much worse since their accidents, that their potential earning powers had been substantially reduced, and that they deserved all the money that the courts would allow them. It was explained that their psychological test results would help determine how large their settlements would be. They were encouraged to fake the most severe disabilities that they could, without making it obvious to the examiner that they were faking. They knew that it was necessary for the technicians to think that they were real clinical patients. The subjects were told nothing about the test battery beyond what is usually told to real patients—that they would be tested for a full working day; that the tests cover a variety of sensory, motor, and cognitive functions; and that the tests measure disabilities that result from brain injuries. They were also given some of the other background information that real trauma patients have: (a) a story about the accident, duration of coma and hospitalization, whether there was a skull fracture, whether seizures developed and, if so, what the seizures were like and how they were being treated (this infor-

² We are grateful for the assistance of Elgan Baker, Gordon Chelune, Charles Clelland, Igor Grant, Robert Ivnik, Charles Matthews, Homer Reed, and James Reed, who served as clinical judges.

³ We appreciate the permission granted by Neil Hillyard and Gerald McDermott to have malingering subjects scheduled for testing under their names.

mation was taken from the files of real head-trauma patients) and (b) a very general description of the neurological lab tests and clinical exams that they would have received if they had been real trauma patients.

The examiner technicians questioned whether 7 of the 16 malingerers had put forth optimal effort on one or more tests. The other 9 malingerers were successful in convincing the examiners that they had given their best performances during all of the testing.

Each test protocol sent to the neuropsychologist judges contained the following: (a) the subject's age, sex, years of education, present/most recent occupation, and time elapsed since the subject's real or pretended head injury; (b) the 3 IQ values and 11 subtest scaled scores on the WAIS; (c) the Halstead Impairment Index and scores on the Category Test (total error score, plus scores on the seven individual groups), Trail Making Test (time and error scores), Tactual Performance Test (times and numbers of blocks placed during all three trials plus Memory and Location scores), Speech-Sounds Perception and Seashore Rhythm Tests, Lateral Dominance Examination, and for each hand on the Finger Oscillation, hand dynamometer, Grooved Pegboard, and Static Steadiness tests; (d) xeroxed data sheets with individual responses recorded on the Aphasia Screening Test and Sensory-Perceptual Examination; and (e) *T* scores for the 3 validity scales and 10 standard clinical scales of the MMPI.

Judges were told of the general design of the study and were provided details of the subject selection procedures and instructions given to the paid malingerers. They knew that some of the subjects were malingerers and that some were non-malingering head-injury patients, but they were not told how many of each were in the total group. Judges were asked to consider each protocol, one at a time, and to make their judgments about the protocol without reference to the others. Furthermore, each judge was asked to consider the 32 protocols in a different random order. After reviewing each protocol, two decisions were required. The first was whether the protocol was probably produced by a malingerer or by a genuine head-injured patient. Second, the degree of confidence with which this first judgment was made was to be rated according to a 4-point scale: very sure, sure, fairly sure, unsure.

Results⁴

Table 1 presents the means, standard deviations, and *t*-test results for the malingering versus head-injury group comparisons on all test measures. The malingerers did as badly as the real head-injury patients in overall level of ability test performance; that is, there were no significant differences between the groups on the three WAIS IQ

values or on the two neuropsychological summary measures (Average Impairment Rating and Halstead Impairment Index). Also, the mean Full Scale IQ obtained by the malingerers when asked to fake was 17 points lower than their actual mean IQ as estimated with the Shipley-Hartford scale, $t(30) = 6.39$, $p < .001$. These results suggest that the malingering group did in fact fake deficits during neuropsychological testing.

Although there were minimal group differences in overall level of performance, the malingerers did differ from the real head-injury group in the pattern of strengths and deficits shown on testing. The real head-injury group did significantly worse than the malingerers on the Category Test; Part B of the Trail Making Test (error component); and on the Total Time, Memory, and Location components of the Tactual Performance Test. The malingerers did worse on the Speech-Sounds Perception Test, the Finger Oscillation Test, finger agnosia, sensory suppressions, hand dynamometer, and WAIS Digit Span. In addition, the malingerers showed (faked) more emotional disturbance on the *F* scale and on six clinical scales of the MMPI.

The neuropsychologist judges correctly classified from 50.0% to 68.8% of the subjects in this study. Sensitivity, or true positive rate for real head injuries, ranged from 43.8% to 81.3%. Specificity, or true negative rate for malingerers, ranged from 25.0% to 81.3%. Three somewhat different measures of diagnostic accuracy are more directly relevant to the clinical situation: the probability that the subject has a real head injury if the judge says he/she does; the probability that the subject is faking if the judge says he/she is; and the judge's overall efficiency, or total correct classification rate for the combined population of head-injury patients and malingerers. These three measures vary not only with the sensitivity and specificity of the judge but also with the prevalence of malingerers in the population being studied. In this study the prevalence rate of malingering was set arbitrarily at 50%, but in general

⁴ Gary Zerbe's assistance with some of the data analyses is gratefully acknowledged.

Table 1

Neuropsychological and Personality Test Means, Standard Deviations, and t-Test Results for Head Injury - Malingering Group Comparisons

Variable	Malingers		Head injuries		t
	M	SD	M	SD	
Neuropsychological summary measure					
Average impairment rating ^a	1.91	.6	1.97	.7	.28
Halstead Impairment Index	.59	.2	.68	.3	1.04
WAIS					
IQ					
Full Scale	96.2	9.7	97.2	11.5	.27
Verbal	98.1	10.3	100.4	10.7	.64
Performance	94.1	11.6	93.1	14.6	.23
Scaled score					
Information	10.1	2.5	10.2	2.3	.15
Comprehension	9.6	3.1	10.1	2.9	.53
Arithmetic	9.2	2.5	9.8	2.6	.56
Similarities	10.6	1.5	10.6	2.7	.08
Digit Span	7.0	2.9	9.5	2.5	2.64*
Vocabulary	10.3	1.1	10.2	1.9	.11
Digit Symbol	6.7	2.5	7.2	3.0	.51
Picture Completion	9.0	2.1	10.0	2.0	1.37
Block Design	10.4	2.6	9.0	2.9	1.48
Picture Arrangement	9.2	2.1	8.1	3.2	1.23
Object Assembly	9.4	2.6	9.7	3.2	.24
Halstead-Reitan battery					
Category Test (errors)	46.1	20.1	67.4	27.2	2.52*
Trail Making Test					
Part A (sec)	55.4	25.2	49.8	32.3	.54
Errors	.12	.3	.31	.5	1.28
Part B (sec)	109.3	54.3	140.8	72.1	1.40
Errors	.81	1.0	2.00	1.9	2.19*
Tactual Performance Test					
Total time/block (min)	.5	.4	1.2	1.3	2.23*
Memory	8.1	1.1	6.2	1.7	3.73***
Location	5.2	2.1	2.8	1.8	3.48**
Speech Sounds Perception (errors)	23.8	12.6	10.6	7.3	3.64***
Seashore Rhythm Test (correct)	21.4	4.5	23.8	4.1	1.60
Finger Oscillation Test ^b					
No./20 sec	63.1	17.1	80.2	21.4	2.49*
Tactile Form Recognition ^b					
Time (sec)	37.6	12.8	32.5	20.8	.83
Finger agnosia (errors) ^b	7.2	5.6	3.5	3.7	2.23*
Finger Tip Writing (errors) ^b	6.5	4.9	5.9	7.3	.26
Suppressions (number) ^b	10.6	7.4	4.1	5.8	2.80**
Crosses (rating) ^a	3.1	1.0	2.7	.9	1.09
Aphasia (errors) ^a	11.2	11.9	9.7	6.1	.45
Added motor tests					
Hand dynamometer (kg) ^b	45.8	20.8	76.4	30.5	3.32**
Grooved Pegboard—time/peg (sec) ^b	3.9	.9	6.3	7.3	1.30
Hole-type steadiness ^b					
Sec	10.1	11.6	26.5	39.3	1.60
Hits	75.3	60.3	99.1	84.0	.92

Table 1—(continued)

Variable	Malingers		Head-injuries		<i>t</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
MMPI (<i>T</i> scores)					
Validity scale					
Lie	49.4	8.2	52.4	10.7	.91
Validity	79.9	18.9	64.1	12.2	2.81**
Defensiveness	48.3	7.4	53.6	9.7	1.73
Clinical scale					
Hypochondriasis	78.3	15.4	60.9	13.8	3.37**
Depression	85.2	15.1	73.2	18.1	2.04
Hysteria	73.6	9.0	61.7	10.6	3.43**
Psychopathic Deviate	69.1	13.7	67.9	10.7	.27
Masculinity-Femininity	55.9	15.9	61.3	9.6	1.16
Paranoia	71.4	13.9	61.1	11.9	2.26*
Psychasthenia	84.2	13.0	66.8	18.0	3.15**
Schizophrenia	93.9	21.2	73.6	22.4	2.64*
Mania	63.5	10.0	65.1	10.0	.44
Social Introversion	73.3	11.2	55.4	10.1	4.74***

Note. *n* = 32 (16 malingers and 16 head-injury patients). WAIS = Wechsler Adult Intelligence Scale; MMPI = Minnesota Multiphasic Personality Inventory.

* Ratings are defined in Russell, Neuringer, and Goldstein (1970).

^b Scores were summed for both sides of the body.

* *p* < .05.

** *p* < .01.

*** *p* < .001.

clinical or litigating populations the prevalence may be much different. Table 2 presents the ranges of these three diagnostic accuracy figures estimated for our 10 judges at three prevalence rates of malingering: 30%, 50%, and 70%. The computations involved are described in Galen and Gambino (1975). In general, the success of the judges

ranged from chance-level prediction to prediction rates about 20% better than chance. For most of the judges, sensitivity exceeded specificity; these judges would do best under the lowest malingering prevalence conditions. Conversely, in this study two judges correctly classified more malingers than head-injury patients, and they would do better

Table 2

Ranges of Diagnostic Accuracies Expected From 10 Neuropsychologists Adjusted for Prevalence of Malingering at Three Levels

Rate	<i>P</i> (I/+)	<i>P</i> (M/-)	Efficiency	Efficiency change from <i>P</i> (M) = .30
<i>P</i> (M) = .30				
Judges' range	.70-.85	.30-.50	.55-.70	—
Chance	.70	.30	.50	—
<i>P</i> (M) = .50				
Judges' range	.50-.70	.50-.69	.50-.69	-.10 to +.08
Chance	.50	.50	.50	.00
<i>P</i> (M) = .70				
Judges' range	.27-.50	.62-.85	.40-.70	-.20 to +.15
Chance	.30	.70	.50	.00

Note. *P*(M) = prevalence of malingers in the population being studied; *P*(I/+) = probability that the patient has a real head injury, if the judge says he/she does; *P*(M/-) = probability that the patient is faking if the judge says he/she is; efficiency = overall proportion of correct classifications to be expected from the judge, given a specified *P*(M).

whenever the prevalence of malingering is higher. The far-right column of Table 2 shows how changing prevalence rates can affect judges' overall efficiencies.

The relationship between the judges' total correct classification rates in this study and their amounts of previous experience in neuropsychology was not statistically significant, $r(8) = .27$. The median percentage of agreement for all possible pairs of judges was only 56 (range = 31%–75%). The relationship between confidence ratings (unsure to very sure) and prediction accuracy was computed for each judge using a point-biserial correlation; that is, confidence ratings for correct versus incorrect judgments were compared. These correlations ranged from $-.13$ to $.46$ ($df = 30$). There was a nonsignificant tendency for more experienced judges to more accurately assess their likelihood of being correct in their subject classifications; a product-moment correlation of $.24$ ($df = 8$) was obtained between judges' years of experience in neuropsychology and their point-biserial correlations for confidence ratings versus classification accuracy. The order in which cases were reviewed did not affect the probability of the cases being correctly identified by the judges, $r(30) = -.20$. Also, there was a nonsignificant correlation between subjects' severity of "impairment" on the neuropsychological test battery (Average Impairment Rating) and subjects' likelihood of being correctly identified by the judges, $r(30) = .17$.

Because the head-injury and malingering subjects did show different patterns of performance on testing, it seemed possible that these differences might permit classification of subjects with a greater degree of accuracy than that achieved by the judges. Two stepwise discriminant function analyses (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975) were performed, one using the neuropsychological variables and the other using the MMPI variables listed in Table 1. The optimal neuropsychological discriminant function cutoff correctly classified all subjects in both groups. The optimal MMPI function cutoff missed only one subject in each group.⁵

It was not possible to recruit more malingerers for direct cross-validation of the

discriminant functions. However, a number of head-injury patients not included in the function development had received clinical evaluations in our neuropsychology laboratory, and their results were available for indirect cross-validation of the functions. Of these patients, 12 had complete neuropsychological data only, 14 had complete MMPI results only, and 62 had both types of data complete. Forty-one of these 88 patients were known to be involved in civil or criminal court actions at the times of testing, and 5 litigators and 1 nonlitigator were considered by the examiners to have been obviously faking on the tests. Thus, 42 of the head-injury patients had reason to exaggerate pathology and/or gave strong clinical evidence of doing so. Another 42 were known not to be involved in court cases and were rated by the examiners as having put forth adequate effort on testing. Excluded from consideration were 2 patients for whom it was uncertain whether they were litigating and 2 who had had cerebrovascular accidents in addition to head injuries.

The malingering discriminant function formulas were applied to the remaining 84 patients. Of the 42 patients who were involved in court cases or who had given strong clinical evidence of exaggerating deficits, 27 (64.3%) were classified as being malingerers by one or both formulas. Of the 42 who had no obvious evidence of or reason to exaggerate pathology, only 11 (26.2%) were classified as malingerers by one or both formulas. The resulting chi-square of 12.95 ($df = 1$, $p < .001$) indicates that obvious fakers and patients involved in civil or criminal court proceedings were significantly more likely to be called malingerers by the formulas.

Discussion

It is difficult to guess what proportion of clinical patients exaggerate deficits on neuropsychological testing. However, the question of exaggeration probably should be considered more often than it is, particularly when there are obvious reasons for the patient not

⁵ Due to the length of these discriminant function formulas, they are not given here. They can be obtained from the first author.

to do his or her best. It is likely that the prevalence of malingering is especially high among patients involved in compensation litigation or competency disputes in criminal trials. This group comprises half of all head-injury referrals to our laboratory. In addition, there are probably other social and personality factors that cause some patients to fake or exaggerate disability. For example, for some individuals the expectation of special care or attention in the family, greater concern from the treating physician, or release from performance expectations might serve as strong inducement for emphasizing real or imagined deficits.

The overall correct classification rates obtained by the neuropsychologist judges in this study are rather modest, but it is unlikely that any of them have had previous experience with known malingerers. It would seem that the judges' prior experience was only indirectly relevant to the questions asked in this study, and this may explain the low correlations between amount of experience and accuracy of judgments. Furthermore, judges were deprived of such potentially useful information as details of the patients' injuries, findings of neurological clinical and laboratory tests, how the recovery periods had gone, how well the patients had done in life prior to their injuries, behavioral observations made by the neuropsychological technicians, and even some test data (verbatim responses on the WAIS). All of this information is usually available in the clinical situation and might contribute to the diagnostic accuracy of the clinician. However, the primary concern of the present study was the value of the test scores in making clinical judgments, and our judges were provided with the same information used in previous studies of clinical interpretation with the Halstead-Reitan battery (Filskov & Goldstein, 1974; Reitan, 1964).

Our group comparisons on the neuropsychological test battery reveal that malingerers can show significant abnormalities on testing, but that the patterns of their strengths and deficits differ from those produced by genuine head-injury patients. These malingerers did especially poorly on motor and sensory tests, but they did relatively well on

several of the cognitive tests that are most sensitive to brain damage. The malingerers also displayed a greater range and degree of apparent personality disturbance on the MMPI, and they tended to obtain high scores on the Validity (*F*) scale. This suggests that the MMPI, in combination with the neuropsychological test battery, has some use in identifying patients who have general tendencies to feign or exaggerate symptoms. The clinician should be doubly wary when a litigation patient shows a "suspicious" pattern of neuropsychological test scores (one that is unusual from a neurological point of view or that resembles the pattern shown by known malingerers) and also gives an MMPI profile of questionable validity.

The results of our discriminant function analyses suggest that group differences in patterns of neuropsychological and MMPI scores are sufficiently reliable to be used in predicting group membership. However, in considering the discriminant functions generated in this study, it is emphasized that any value they may have is restricted to the head injury-malingerer distinction. Two patients who had suffered strokes as well as head injuries serve as good examples of this. They were both classified as malingerers by the neuropsychology function, due to severe sensory-motor deficits, which are much more typical of cerebrovascular accidents than of traumatic head injuries.

The need to cross-validate or improve our discriminant functions is also emphasized, particularly in view of the small subject groups on which they were developed. In such research, the design used in the present study could be extended by including cooperative, nonlitigating head-injured patients who are asked to make themselves look as impaired as possible on the testing. In the clinical situation, many malingerers will have some real deficits to "build on." This may make them more difficult to identify than were our normal malingerers, and the suggested extension of the present study would therefore be of value. Nevertheless, the results of our indirect cross-validation of the discriminant functions suggest that these functions probably do have some reliability

in detecting those real head-injury patients who exaggerate their pathology.

On virtually all ability tests, the subject is told what is required in order to do well. At the same time, it usually becomes obvious what a bad performance entails, for example, be slow, make errors, fail to solve the problem. Therefore, neuropsychological tests would seem intrinsically vulnerable to faking. Faking certain "objective" aspects of the physical neurological exam (muscle wasting, asymmetrical or pathological reflexes, nystagmus) would be more difficult, but sensory and mental status testing by the neurologist should be just as easily faked as are neuropsychological tests. Clear discrepancies between objective and subjective parts of the neurological exam may give the malingerer away, or at least arouse suspicion. However, there may be no objective findings, and the rather "spotty" (mild and inconsistently localized) symptoms presented by many patients with histories of head injuries make the real versus faking distinction much more difficult. In cases in which this judgment has to be made primarily on the basis of fakable symptoms and complaints of mental disturbance, any patterns of such symptoms that are characteristic of malingerers probably would be more reliably identified by standardized and comprehensive neuropsychological testing than by nonstandardized and briefer clinical examinations. Nevertheless, until more is known about fakers' patterns of performance, caution is warranted in interpreting test results of patients who may have reason to exaggerate pathology. It is also possible that malingerers who have been expertly coached will be able to simulate more successfully the deficits of genuinely brain-damaged patients. Under these circumstances discriminant functions of the type developed here will be less reliable in assigning patients to the nonmalingering category.

References

- Anthony, N. C. Comparison of clients' standard, exaggerated and matching MMPI profiles. *Journal of Consulting and Clinical Psychology*, 1971, 36, 100-103.
- Boll, T. J., Heaton, R. K., & Reitan, R. M. Neuropsychological and emotional correlates of Huntington's chorea. *Journal of Nervous and Mental Disease*, 1974, 158, 61-69.
- Cofer, C. N., Chance, J. E., & Judson, A. J. A study of malingering on the MMPI. *Journal of Psychology*, 1949, 27, 491-499.
- Dikmen, S., & Reitan, R. M. MMPI correlates of adaptive ability deficits in patients with brain lesions. *Journal of Nervous and Mental Disease*, 1977, 165, 247-254.
- Filskov, S. B., & Goldstein, S. G. Diagnostic validity of the Halstead-Reitan neuropsychological battery. *Journal of Consulting and Clinical Psychology*, 1974, 42, 382-388.
- Galen, R. S., & Gambino, S. R. *Beyond normality: The predictive value of medical diagnoses*. New York: Wiley, 1975.
- Gough, H. G. Simulated patterns on the MMPI. *Journal of Abnormal and Social Psychology*, 1947, 42, 215-225.
- Gough, H. G. The F minus K dissimulation index for the MMPI. *Journal of Consulting Psychology*, 1950, 14, 408-413.
- Gough, H. G. Some common misconceptions about neuroticism. *Journal of Consulting Psychology*, 1954, 18, 287-292.
- Kløve, H. Validation studies in adult clinical neuropsychology. In R. M. Reitan & L. A. Davison (Eds.), *Clinical neuropsychology: Current status and applications*. New York: Wiley, 1974.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. *SPSS: Statistical package for the social sciences* (2nd ed.). New York: McGraw-Hill, 1975.
- Osborne, D. A. A moderator variable approach to MMPI validity. *Journal of Clinical Psychology*, 1970, 26, 486-490.
- Paulson, M. J., & Lin, T. T. Predicting WAIS I. Q. from Shipley-Hartford scores. *Journal of Clinical Psychology*, 1970, 26, 453-461.
- Reitan, R. M. Psychological deficits resulting from cerebral lesions in man. In J. M. Warren & K. A. Akert (Eds.), *The frontal granular cortex and behavior*. New York: McGraw-Hill, 1964.
- Reitan, R. M. A research program on the psychological effects of brain lesions in human beings. In N. R. Ellis (Ed.), *International review of research in mental retardation* (Vol. 1). New York: Academic Press, 1966.
- Reitan, R. M., & Davison, L. A. (Eds.). *Clinical neuropsychology: Current status and applications*. New York: Wiley, 1974.
- Russell, E. W., Neuringer, C., & Goldstein, G. *Assessment of brain damage: A neuropsychological key approach*. New York: Wiley, 1970.
- Shaw, D. J., & Matthews, C. G. Differential MMPI performance of brain-damaged versus pseudo-neurologic groups. *Journal of Clinical Psychology*, 1965, 21, 405-408.
- Shipley, W. C. A self-administering scale for measuring intellectual impairment and deterioration. *Journal of Psychology*, 1940, 9, 371-377.
- Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation, 1955.

Validity of Self-Reports in Three Populations of Alcoholics

Linda C. Sobell

Dede Wallace Center, Nashville, Tennessee and
Vanderbilt University

Mark B. Sobell

Vanderbilt University

This study examined whether population type (voluntary outpatient, voluntary inpatient, coerced outpatient) and question type (alcohol, nonalcohol, demographic) differentially affected the validity of alcoholics' self-reports. Three distinctly different populations of alcoholics independently completed life history questionnaires. The veridicality of subjects' answers was assessed using official records and documents. Generally, alcoholics in this study gave highly valid self-reports, a result that parallels the findings of earlier studies. Question type differentially affected the validity of subjects' interview answers, as significantly fewer invalid answers were given to demographic questions than to alcohol and nonalcohol questions. Population type, however, did not significantly affect the validity of self-reported life history information. Invalid interview answers were more often overreported than underreported when compared with official records.

Despite limited empirical evidence, skepticism about the validity of alcoholics' self-reports has abounded (reviewed in Sobell, 1976). Yet, in the alcoholism field, self-reports are a major source of data. From a practical standpoint, self-reports are convenient and economical, they obviate checking official record sources, and they often provide information when further verification is impossible. Given their widespread use, it seems highly unlikely that self-reports will be abandoned as a primary source of information.

Although the alcoholism field has long relied on self-reported data, however, only recently has the validity of alcoholics' self-reports been examined (Armor, Polich, & Stambul, 1976; Sobell, 1976; Sobell & Sobell, 1975; Sobell, Sobell, & Samuels, 1974; Sobell, Sobell, & VanderSpek, Note 1). These investigations have found that most verifiable self-

reported life history data by alcoholics are generally quite valid. However, beyond consideration of the overall validity of alcoholics' self-reports, little is known about the validity of subjects' answers as a function of population type (i.e., outpatients vs. inpatients, females vs. males, chronics vs. problem drinkers, voluntary vs. coerced, etc.). Will all alcoholics be equally self-disclosing, cooperative, and, moreover, truthful in answering different types of questions? For example, chronic alcoholics might have more difficulty in accurately remembering certain events simply because they may have more events to recall than other populations of alcoholics. Similarly, alcoholics coerced into treatment might be relatively less self-disclosing about their drinking behavior if they anticipate that a more severe drinking history is suggestive of a need for more treatment.

The present study examined whether the validity of self-reports differs between and within three different populations of alcoholics. The investigation was limited to examining self-reports that could be verified by official records.

This article was presented, in part, at the 22nd annual meeting of the Southeastern Psychological Association, New Orleans, Louisiana, March 1976.

Requests for reprints should be sent to L. Sobell, Dede Wallace Center, P. O. Box 40487, Nashville, Tennessee 37204.

Method

Subjects

Three distinctly different groups of male alcoholics served as subjects in this study: (a) voluntary outpatient alcoholics (V/OP) who were clients at the Orange County, California, Department of Mental Health Alcoholism Services (OCAS; $n=14$); (b) coerced outpatient alcoholics (C/OP) who were court referred for treatment at the OCAS ($n=12$); and (c) voluntary inpatient alcoholics (V/IP) who were hospitalized at the San Diego County, California, Detoxification Center ($n=13$). Subjects were further selected in accordance with the following criteria: (a) no evidence of alcohol withdrawal symptoms or alcohol intoxication at the time of the interview; (b) no evidence of organic brain syndromes or a primary diagnosis other than alcoholism; and (c) voluntary participation in the study. No subject cleared for the study refused to participate.

Since the three groups of subjects were specifically chosen to differ from one another, statistical analyses comparing descriptive characteristics between groups of subjects were not performed. However, demographic variables clearly reflected differences between the three groups. Subjects in the V/IP group were older than subjects in the other two groups (M age: V/IP = 44.5 years, V/OP = 38.4 years, C/OP = 34.2 years), reported longer drinking problem histories (M years: V/IP = 16.8, $n=12$; V/OP = 8.6; C/OP = 3.6), and had more alcohol-related arrests (M arrests: V/IP = 19.5, V/OP = 4.5, C/OP = 4.5). Although few subjects reported alcohol-related hospitalizations, subjects in the V/OP group reported a mean of .79 such hospitalizations compared to a mean of .38 hospitalizations for Group V/IP. No subject in Group C/OP reported any alcohol-related hospitalizations. Subjects in Group V/IP also had about three times as many nonalcohol arrests as subjects in the other two groups (M nonalcohol arrests: V/IP = 4.0, C/OP = 1.3, V/OP = 1.2). The largest observed difference among groups was in terms of ethnicity. Even though two of the groups had similar percentages of Caucasian subjects (V/IP = 92.3%, V/OP = 92.9%), only 50.0% of C/OP subjects were Caucasian. Finally, all groups of subjects had a mean education of about 12 years.

Procedure

Group interviews were conducted separately with subjects in each of the three experimental groups. In the group setting, each subject was asked to complete a questionnaire and return it to the investigator (L. Sobell) when finished. The questionnaire contained 35 verifiable questions about drinking and life history information.

At the time of the interview, subjects were not aware that their answers would be compared with official records. However, after the questionnaire was completed, all subjects were debriefed as to the

nature of the study. Subjects were also assured that their interview answers would be confidential, would not be recorded in their clinic chart, and would not influence the kind of treatment received. Coerced subjects were further assured that participation in the study would not affect their legal status nor would the court or the treatment agency have access to their answers.

Official Record Sources

Interview answers were compared with the following official records: (a) driver records from the California Department of Motor Vehicles; (b) official arrest records. [Commonly known as "rap sheets," these records were those of the California Bureau of Criminal Identification and Investigation (CII) and the Federal Bureau of Investigation (FBI). In California, all legal agencies must furnish the Bureau of CII with daily reports of virtually all misdemeanor and felony arrests within their jurisdiction and the eventual disposition of such charges. FBI records contain essentially the same information but on a national basis. The information provided by these records constituted a known minimum number of arrests for any given individual.]; (c) inpatient hospitalizations at Orange County, California, Medical Center, the sole public hospital serving the county (for Groups V/OP and C/OP only); (d) inpatient hospitalizations at San Diego County, California, Detoxification Center—the sole public detoxification facility in San Diego County (for Group V/IP only); (e) inpatient hospitalizations at Metropolitan State Hospital (Norwalk, California), the state hospital serving both Orange and San Diego Counties; and (f) inpatient hospitalizations at all other California state hospitals prior to July 1969—these hospitalizations were listed on the CII record until July 1969.

Dependent Measures

Each of the 35 verifiable interview questions was grouped into one of three mutually exclusive categories: (a) *alcohol questions* ($n=7$)—questions about alcohol-related behaviors; (b) *nonalcohol questions* ($n=19$)—questions about behaviors not directly related to drinking; and (c) *demographic questions* ($n=9$)—questions about personal identifying information (i.e., name, date of birth, age, hair color, social security number, etc.). To be scored as valid, interview answers had to be identical with the record data. Invalid answers to all alcohol questions and 16 of the 19 nonalcohol questions were further evaluated as either (a) *overreports* (information that was reported in the interview but that did not appear on official records) or (b) *underreports* (information that was not reported in the interview but that did appear on official records).

The questionnaire also included two questions to evaluate whether subjects who responded affirmatively to fictitious questions would also tend to over-

report other aspects of their behavior. The first question asked subjects if they had ever used a drug called "bindro." This fictitious drug (previously used in a study by Petzel, Johnson, & McKillip, 1973) was disguised among a list of 11 other "real" drugs. The second fictitious question asked subjects how many times they had been hospitalized for alcohol-related problems at "Garden View Rehabilitation Hospital." This question, too, was disguised among a series of questions about admissions to actual hospitals for alcohol-related problems.

Results

Nonresponse Bias

Nonresponse bias, the failure of subjects to respond to questions, was not an important factor in this study. All subjects in Groups V/OP and V/IP answered all interview questions, whereas subjects in Group C/OP answered 98.7% of all interview questions. (Two subjects each failed to answer one question, and one subject failed to answer four questions.)

Validity of Interview Answers

The two groups of alcoholics who were in treatment voluntarily, Groups V/OP and V/IP, validly answered 87.8% and 80.0%, respectively, of all verifiable questions. Alcoholics coerced to participate in treatment (Group C/OP) validly answered 83.6% of all verifiable questions. The overall percentages of invalid answers for each type of question for each of the three groups of subjects were (a) for Group V/OP, alcohol = 18.4%, nonalcohol = 11.3%, and demographic = 9.5%; (b) for Group C/OP, alcohol = 9.5%, nonalcohol = 20.3%, and demographic = 13.9%; and (c) for Group V/IP, alcohol = 25.3%, nonalcohol = 22.7%, and demographic = 10.3%. Only two subjects responded positively to the fictitious questions. Both subjects were in Group V/IP, and each reported having been hospitalized one time at Garden View Rehabilitation Hospital.

The hypothesis that population type and question type would differentially affect the validity of subjects' interview questions was tested using a 3×3 repeated measures analysis of variance. Population type was treated as an independent factor with three levels

(V/OP, C/OP, V/IP), and question type was treated as a dependent factor with three levels (alcohol, nonalcohol, and demographic). Since there were unequal numbers of the three types of questions, subjects' raw scores for number of invalid answers were converted into proportions of invalid answers for each of the three question types. Questions that were not answered (missing data) or reported by subjects as "unknown" were scored as invalid. A few subjects answered some questions by giving a range (e.g., 4-7 public drunk arrests). These answers were converted to a single value by obtaining the arithmetic mean of the range. Consequently, not all interview answers were integers. Interview answers that were given as a range were not scored as invalid if they were within $\pm .5$ of the official record information.

The analysis of variance indicated a significant main effect for question type, $F(2, 52) = 3.61$, $p < .05$, with no significant main effect for population type and no significant interaction. A simple effects analysis was performed to explore the significant main effect for question type. A significant simple effect, $F(2, 52) = 21.67$, $p < .01$, indicated that fewer invalid answers were given to demographic questions ($M = 4.3\%$) than to alcohol ($M = 7.0\%$) and nonalcohol ($M = 7.0\%$) questions. This finding was not surprising, as most demographic questions required answers that were personally descriptive and fixed (e.g., name, date of birth, social security number, eye color, etc.). The two demographic questions answered least validly across all subjects were hair color and height. The answers to these questions were compared with data taken from the subjects' driver records. Since these two variables were almost susceptible to real change over time, higher invalidity on these variables could reflect actual change.

Direction of Discrepancy

Demographic questions were excluded from analysis of direction of discrepancy, because invalid answers for eight of the nine questions could not be categorized as either overreports or underreports (e.g., date of birth, ethnicity, eye color, etc.). Three nonalcohol

Table 1
Percentage of Subjects Who Gave Valid, Overreported, and Underreported Interview Answers As Compared With Official Records

Interview question (condensed)	V/OP (n = 14)			C/OP (n = 12)			V/IP (n = 13)		
	% valid	% over-reported	% under-reported	% valid	% over-reported	% under-reported	% valid	% over-reported	% under-reported
Alcohol questions									
No. hospitalizations									
OCMC	78.57	7.14	14.29	91.67	8.33	.00	69.23	.00	30.77
MSH	92.86	.00	7.14	100.00	.00	.00	100.00	.00	.00
CSH	78.57	21.43	.00	100.00	.00	.00	61.54	23.08	15.38
No. arrests									
Public drunk	92.86	7.14	.00	100.00	.00	.00	100.00	.00	.00
Drunk in auto	100.00	.00	.00	100.00	.00	.00	61.54	38.46	.00
Open container	100.00	.00	.00	100.00	.00	.00	76.92	23.08	.00
No. drunk driving convictions	42.86	50.00	7.14	41.67	50.00	8.33	53.85	30.77	15.38
Nonalcohol questions									
No. arrests									
Assault & battery	100.00	.00	.00	100.00	.00	.00	69.23	15.38	15.38
Other crimes	92.86	7.14	.00	63.64*	.00*	36.36*	23.08	15.38	61.54
No. convictions									
Driver's license suspended/ revoked Calif. driver's license	100.00	.00	.00	83.33	8.33	8.33	84.62	15.38	.00
Reckless driving	57.14	42.86	.00	58.33	33.33	8.33	61.65	38.46	.00
No. speeding tickets in Calif. last 2 yr.	64.29	35.71	.00	75.00	16.67	8.33	92.31	7.69	.00
No. times Calif. driver's license was									
Revoked	92.86	7.14	.00	91.67	.00	8.33	84.62	7.69	7.69
Suspended	85.71	14.29	.00	50.00	50.00	.00	53.85	30.77	15.38
No. auto accidents in last 3 yr.	85.71	7.14	7.14	63.64*	18.18*	18.18*	84.62	15.38	.00

Table 1 (continued)

Interview question (condensed)	V/OP (n = 14)			C/OP (n = 12)			V/IP (n = 13)		
	% valid	% over-reported	% under-reported	% valid	% over-reported	% under-reported	% valid	% over-reported	% under-reported
No. arrests									
Drugs	100.00	.00	.00	91.67	8.33	.00	100.00	.00	.00
Burglary	92.86	.00	7.14	75.00	8.33	16.67	92.81	7.69	.00
Petty theft	92.86	.00	7.14	100.00	.00	.00	84.62	7.69	7.69
Robbery	92.86	.00	7.14	100.00	.00	.00	84.62	7.69	7.69
Assault with a deadly weapon	100.00	.00	.00	91.67	.00	8.33	84.62	7.69	7.69
Disturbing peace	78.57	14.29	7.14	91.67	8.33	.00	53.85	46.15	.00
No. times on									
Probation	57.14	7.14	35.71	54.55 ^a	18.18 ^a	27.27 ^a	38.46	23.08	38.46
Parole	100.00	.00	.00	90.91 ^a	9.09 ^a	.00 ^a	100.00	.00	.00
Currently on									
Probation ^b	92.86	.00	7.14	75.00	25.00	.00	84.62	15.38	.00
Parole ^b	100.00	.00	.00	100.00 ^a	.00 ^a	.00 ^a	100.00	.00	.00
Valid Calif. driver's licence	100.00	.00	.00	91.67	8.33	.00	92.31	7.69	.00

Note. V/OP = voluntary outpatient alcoholics; C/OP = coerced outpatient alcoholics; V/IP = voluntary inpatient alcoholics; OCMC = Orange County Medical Center (for V/IP San Diego Detox Center); MSH = Metropolitan State Hospital; CSH = California state hospitals.

^a N = 11.

^b Answers for these questions required only a yes or no response.

questions were also excluded because they only required a yes or no answer. Further, alcohol and nonalcohol questions that were not answered (missing data) or subjects who reported unknown answers were not included in these analyses because direction of discrepancy could not be determined.

Table 1 presents the percentage of valid, overreported, and underreported interview answers given by each of the three groups of subjects for alcohol questions and nonalcohol questions. The percentages of valid answers for combined alcohol and nonalcohol questions ($n = 26$) were very similar for both outpatient groups (C/OP = 84.0%, V/OP = 86.8%). Of the invalid answers given by subjects in each of these two groups, approximately two thirds were overreported. Although the data for the inpatient group of subjects (V/IP) followed the same proportional distribution of overreports and underreports as the outpatient groups, the absolute percentage of overreports and underreports was about twice that of the outpatient subjects. Subjects in the inpatient group gave the lowest percentage of valid answers (68.0%) to questions for which direction of discrepancy could be determined.

One-sample t tests (two-tailed) were performed for each group of subjects to determine if the mean proportion of overreported answers differed significantly from chance ($p = .5$) as compared to underreported answers. The mean proportions of overreports and underreports for each group were as follows: C/OP—overreports = .65, underreports = .35; V/OP—overreports = .74, underreports = .26; and V/IP—overreports = .56, underreports = .44. For the two outpatient groups, C/OP and V/OP, the mean proportion of overreported answers was significantly greater than chance, $t(11) = 1.93$, $p < .05$, and $t(13) = 4.08$, $p < .05$, respectively. On the other hand, the mean proportion of overreported answers for the inpatient group, V/IP, did not differ significantly from chance, $t(12) = .64$, $p > .05$.

Discussion

The overall validity rates for combined alcohol and nonalcohol questions for both

groups of outpatient alcoholics is comparable to that obtained in other studies (Ball, 1967; Guze, Tuason, Stewart, & Picken, 1963; Sobell & Sobell, 1975; Sobell et al., 1974). In fact, the percentages of valid, overreported, and underreported interview answers for both groups of outpatient alcoholics in this study are nearly identical to results reported by Sobell and Sobell (1975) for a similar population. More importantly, the self-reports of alcoholics who were coerced (court referred) into treatment were found to be as valid as those given by alcoholics who voluntarily entered treatment. The finding that all three groups of subjects gave more overreported than underreported interview answers is also consistent with earlier research (Ball, 1967; Guze et al., 1963; Knupfer, in press; Sobell, 1976; Sobell & Sobell, 1975; Sobell et al., 1974).

In the last few years there has been a proliferation of studies using fictitious drug questions to assess overreporting of drug use. In all of these studies, only negligible percentages of fictitious drug use have been reported [Haberman, Josephson, Zanes, & Ellison, 1972 (1%); Whitehead & Smart, 1972 reviewed two studies reporting less than 1% and one study reporting 7.5%; Petzel et al., 1973 (3.8%); Single, Kandel, & Johnson, 1975 (< 1%)]. Even though no subject in the present experiment reported using the fictitious drug bindro, two V/IP subjects reported being hospitalized at the bogus Garden View Rehabilitation Hospital. Given this paucity of results, the utility of fictitious questions to identify subjects who gave invalid self-reports cannot be evaluated.

Finally, although different populations of alcoholics in this study generally gave quite valid self-reports, this finding cannot necessarily be extrapolated to conclude that alcoholics are basically honest in their daily interactions. For instance, alcoholics who are not interviewed within the context of an alcohol treatment program and who are not guaranteed the confidentiality of their answers might engage in denial, misrepresentation, and/or lying, especially if they do not view themselves as having alcohol-related problems. It is notable, however, that individuals who were court referred and coerced into an outpatient

alcohol treatment program did not give more invalid self-reports than voluntary outpatients or inpatient alcoholics.

Reference Note

1. Sobell, L. C., Sobell, M. B., & VanderSpek, R. *Three independent comparisons between clinical judgment, self-reports and physiological measures of blood alcohol level*. Paper presented at the 22nd annual meeting of the Southeastern Psychological Association, New Orleans, Louisiana, March 1976.

References

- Armor, D. J., Polich, J. M., & Stambul, H. B. *Alcoholism and treatment*. (Report R-1739-NIAAA.) Santa Monica, Calif.: Rand Corporation, 1976.
- Ball, J. C. The reliability and validity of interview data obtained from 59 narcotic drug addicts. *American Journal of Sociology*, 1967, 72, 650-654.
- Guze, S. B., Tuason, V. B., Stewart, M. A., & Picken, B. The drinking history: A comparison of reports by subjects and their relatives. *Quarterly Journal of Studies on Alcohol*, 1963, 24, 249-260.
- Haberman, P. M., Josephson, E., Zanes, A. & Ellison, J. High school drug behavior: A methodological report on pilot studies. In E. Josephson & E. E. Carroll (Eds.), *Student drug surveys*. Farmingdale, N.Y.: Baywood, 1972.
- Knupfer, G. The validity of survey data on drinking problems: A comparison between respondent's self-reports and outside sources of information. In W. J. Clark & G. Knupfer (Eds.), *Drinking patterns and problems. A decade of studies in the San Francisco area*. New Brunswick, N.J.: Rutgers Center of Alcohol Studies, in press.
- Petzel, T. P., Johnson, J. E., & McKillip, J. Response bias in drug surveys. *Journal of Consulting and Clinical Psychology*, 1973, 40, 437-439.
- Single, E., Kandel, D., & Johnson, B. D. Reliability and validity of drug-use responses in a large-scale longitudinal survey. *Journal of Drug Issues*, 1975, 5, 426-443.
- Sobell, L. C. The validity of self-reports: Toward a predictive model. (Doctoral dissertation, University of California, Irvine, 1976). *Dissertation Abstracts International*, 1976, 37, 1469B. (University Microfilms No. 76-19, 632)
- Sobell, L. C., & Sobell, M. B. Outpatient alcoholics give valid self-reports. *Journal of Nervous and Mental Disease*, 1975, 161, 32-42.
- Sobell, M. B., Sobell, L. C., & Samuels, F. H. The validity of self-reports of prior alcohol-related arrests by alcoholics. *Quarterly Journal of Studies on Alcohol*, 1974, 35, 276-280.
- Whitehead, P. C., & Smart, R. G. Validity and reliability of self-reported drug use. *Canadian Journal of Criminology and Corrections*, 1972, 14, 1-7.

Received May 20, 1977 ■

Drug Preferences of Multiple Drug Abusers

Robert J. Harford

Department of Psychiatry, Yale Medical School

The aims of this study were (a) to develop a reliable measure of preferences among types of controlled drugs and (b) to examine the correspondence between the most preferred drug and the drug most frequently used. One hundred thirty active multiple drug abusers rated their preferences among 11 combinations of controlled drugs and common methods of administration using the method of paired comparisons. Edwards' coefficient of consistency indicated that preferences were highly consistent (.92) and therefore internally reliable. Nearly half of the respondents most preferred drugs other than the type that they most frequently used, and their preferences were related to the method of administration. The results suggest that preference is one among several determinants of drug use.

Attempts to identify personality characteristics that predispose drug abusers to continued use of controlled drugs have produced generally inconclusive or conflicting results. Drug-related variables such as physiological addictiveness, cost, and availability as well as social variables such as setting, peer pressure, and acceptability seem to have stronger effects than personality in determining how frequently various drugs are used.

Several recent articles (Crain, Ertel, & Gorman, 1975; Penk & Robinowitz, 1976; Ungerer, Harford, Brown, & Kleber, 1976) have suggested that personality characteristics are related primarily to drug preference rather than actual drug use. Preference among alternative drugs can be conceptualized as an intervening variable that results from interactions among the users' psychological characteristics (including personality, attitudes, and expectations), situational variables, and differences in the psychopharmacological effects

of specific types of drugs. Preference may be one of several factors that determine frequency of use. If preference mediates the relationship between personality and drug-taking behaviors, personality constructs are expected to be more strongly related to drug preference than to frequency of use. This formulation could account for previous inconsistent findings and could contribute to the development of our understanding of the personality dynamics underlying compulsive drug use.

Although several personality correlates of drug preference have been reported, the specific relationships between preference and use have not been examined. Some investigators (Baer & Corroda, 1973; Henriques, Arsenian, Cutter, & Samaraweera, 1972) have assumed that the most frequently used drug is the most preferred. Since there is no evidence to the contrary, the possibility that preference and use are synonymous cannot be discounted, but the hypothesized differential relationships between (a) personality and preference and (b) personality and drug use imply that preference and use are only moderately related. The explanatory value of the drug preference construct is negligible if the measures of preference and use are highly correlated.

One limitation of drug preference research is that reliabilities of the preference measures

This research was supported in part by Grants DA 01097 and DA 16356 from the National Institute on Drug Abuse, Herbert D. Kleber, principal investigator. The results were presented at the Midwestern Psychological Association, Chicago, May 1976.

Requests for reprints should be sent to Robert J. Harford, Department of Psychiatry, Yale University School of Medicine, 98 Park Street, New Haven, Connecticut 06519.

generally are unknown. If preferences among drugs are not well measured, results that seem to confirm the predicted moderate degree of correspondence between preference and use may be the product of error rather than an indication of true differences between constructs. The single-item rating scales typically used to measure preference seem to be particularly susceptible to unreliable responding that could artifactually attenuate the relationship between preference and use.

The present research examined the drug preferences of a group of active multiple drug abusers who were referred for treatment. Drug preference was measured by a paired-comparisons methodology that included six drug types (i.e., amphetamine, barbiturate, cocaine, hallucinogen, cannabis, opiate) and encompassed the three most common methods of drug administration (i.e., oral, intranasal, intravenous). The paired-comparisons method allows a direct test of consistency (Edwards, 1957) that can be interpreted as an index of internal reliability of the preference scores. The relationship between the drug preference construct and drug use was assessed by comparing the most preferred drug with the drug used most frequently during the preceding 60-day period.

Method

Subjects

A drug preference inventory was administered to 130 persons seeking treatment for drug dependence at the Connecticut Mental Health Center in New Haven. Seventy-two percent of the sample were males, and 63% were white. Their ages ranged from 17 to 29, with a mean of 22.5 years.

Current Drug Use

The type of drug used most frequently during the preceding 60 days was determined by self-report and, in most cases, was corroborated by medical examination and/or thin-layer chromatography urinalysis results. For the majority of respondents, the most frequently used drug was heroin (60.0%), followed by marijuana (17.6%), barbiturates (8.4%), amphetamines (7.6%), hallucinogens and psychedelics (5.0%), and other opiates (1.4%).

Drug Preference

The drug preference inventory contains all possible $[n(n-1)/2]$ pairwise comparisons of 11 combinations of types of controlled drugs and commonly used methods of administration: smoking cannabis, snorting opiates, ingesting opiates, shooting opiates, snorting cocaine, shooting cocaine, ingesting amphetamines, shooting amphetamines, ingesting barbiturates, shooting barbiturates, and ingesting hallucinogens. Each of the 55 preference pairs are rated on a 6-point scale, which ranges from -3 to +3. There is no indifference point. The inventory also includes a self-report measure of whether the respondent has ever used each of the 11 drug categories.

Procedure

The testing procedure was presented as voluntary research directed toward helping the clinics improve the quality of treatment. Applicants were guaranteed confidentiality and were assured that their performance would not influence their treatment status. Three applicants declined to participate.

Results

Previous Drug Use

Responses to the ever used items indicated that at some time during their lives, 94% of the respondents had smoked cannabis, 61% had snorted cocaine, 61% had shot cocaine, 70% had ingested opiates, 56% had snorted opiates, 63% had shot opiates, 48% had ingested amphetamines, 77% had shot amphetamines, 66% had ingested hallucinogens, 43% had ingested barbiturates, and 72% had shot barbiturates. Seventy-six percent had used more than 4 of the 11 drug categories, and 44% had used at least 9. The self-reports indicate that the majority of respondents actually had used a variety of controlled drugs, and their preferences were based at least in part on personal experience with their effects.

Consistency of Preferences

The individual preference matrices were tested for transitivity using the coefficient of consistency (zeta). For the 55-item matrix, a ζ of .42 indicates greater than chance consistency at the .05 level of confidence, and a ζ of .64 is significant at the .001 level. The mean ζ (.92) indicated that preferences were

highly consistent and transitive. Seven respondents were excluded from subsequent analyses because their preference data yielded consistency coefficients that were less than .53 ($p > .01$). Two of the excluded respondents had consistency coefficients greater than .42. The higher cutoff score was selected to minimize the possibility that unreliability of preference contributed to any lack of correspondence between preference and use.

Preferred Drugs and Methods of Administration

Preference scores ranging from 0 to 10 were calculated for each of the 11 drug categories by counting the number of times the category was preferred to all others. When the data were grouped within drugs, the most preferred drug was opiates (43%) followed by cocaine (19%), cannabis (19%), amphetamines (8%), hallucinogens (6%), and barbiturates (4%). When the methods of administering the drugs were compared, 56% of the respondents preferred intravenous cocaine to intranasal cocaine, 61% preferred intravenous opiates to intranasal opiates, 59% preferred intravenous opiates to oral opiates, 78% preferred intravenous barbiturates to oral barbiturates, and 53% preferred intravenous amphetamines to oral administration of amphetamines. These results show that for the four injectable drugs, the majority of respondents preferred intravenous to other methods of administration.

Preferences and Use

When the most preferred drug was compared with the currently used drug, 46% of this sample most frequently used a drug other than the one that they most preferred. Forty-five of the 75 opiate users preferred opiates to all other drugs. One of the 11 barbiturate users preferred barbiturates. (Three preferred cocaine, and 4 preferred opiates.) Six of the 10 amphetamine users preferred amphetamines. Twelve of the 21 cannabis users preferred cannabis, and 3 of the 6 hallucinogen users preferred hallucinogens. Thus, 60% of the opiate users preferred opiates. (Thirteen

percent preferred cannabis, and 25% preferred cocaine to opiates.) Forty-six percent of those who most frequently used drugs other than opiates preferred their current drug. (Seventeen percent preferred opiates, 9% preferred cocaine, and 9% preferred hallucinogens.) Cocaine was the drug least often identified as the most frequently used drug, but it was the most frequently preferred alternative to the currently used drug (17.5%), followed by cannabis (8%) and opiates (7%).

Discussion

The method of paired comparisons yielded internally consistent ratings of drug preference. Since zeta measures transitivity, the results demonstrate that drug preferences are transitive. Transitivity implies that, for example, if intravenous cocaine is preferred to intravenous opiates, and intravenous opiates are preferred to smoking cannabis, then intravenous cocaine is preferred to smoking cannabis. The high coefficient of consistency indicates that the method of paired comparisons is a reliable procedure for measuring drug preference. Since reliability of self-reports is particularly problematic among drug abusers in treatment, this finding suggests that the drug preference inventory would produce reliable preference scores in other populations as well.

A scant majority of the drug abusers in this study most frequently used the type of drug that they most preferred. The discrepancy between preference and use was greater among the nonopiate users than among the opiate users. However, nearly 40% of the opiate users preferred either cocaine or marijuana to opiates, whereas only 17% of the nonopiate users preferred opiates to all other drugs.

The results indicate that preference and use are more independent than has been believed previously. The assumption that the most frequently* used drug is the *de facto* drug of choice (Henriques et al., 1972) was not confirmed. Many drug abusers compulsively use a drug other than the one that they most prefer. The reasons for the less than perfect correspondence between preference

and use are not fully evident, but the high coefficients of consistency obtained for the drug preference inventory demonstrate that error in measuring preference was not responsible for attenuating the relationship.

Some of the reasons for using drugs other than the most preferred type may be physiological or economic. For example, when the supply of a nonaddictive preferred drug is limited temporarily, less preferred but physiologically more addictive drugs may be substituted. As a combined result of tolerance and withdrawal effects that induce increased use of addictive drugs, use of the less preferred drug may remain higher than that of the preferred drug after the supply of the preferred drug expands. This process might account for the current addictions of some of the opiate users who continued to prefer non-addictive cocaine and marijuana.

In the dosages used by many street addicts, heroin functions primarily to avoid withdrawal rather than to achieve euphoria. Cocaine is rarely available in quantities sufficient to produce tolerance effects, and tolerance to cannabis does not develop. Consequently, preferences for cocaine and marijuana by opiate users might derive in part from their continued ability to produce euphoria that can no longer be attained with opiates.

The results are consistent with the hypothesis that preference is one among several determinants of drug use, but they do not exclude the possibility of a reciprocal causal relationship between preference and use. Prolonged frequent use of a most preferred drug may lead to its devaluation in the drug preference structure. For individuals whose use of drugs is motivated by sensation seeking (Zuckerman, Bone, Neary, Mangelsdorff, & Brustman, 1972), for example, increased familiarity with the effects of the most frequently used drug may enhance preference for more exotic, less frequently used drugs. Experience with the long-term negative effects of heroin addiction also could be expected to produce increased preference for nonaddictive drugs. Preference may affect use primarily during the acquisition phase of compulsive drug use, whereas greater effects of drug use

on preference could be expected during the maintenance phase. Longitudinal investigations of preference and use in nonaddict populations would be helpful in clarifying the complex interrelationships among these variables.

The results show that within each of the four types of injectable drugs the majority of respondents preferred intravenous to other methods of administration. This finding indicates that preference depends in part on the method of administration, but the reasons for the popularity of the intravenous method remain a matter for speculation. Gossop and Connell's (1975) explanation for the higher evaluation of drugs by oral abusers than by intravenous abusers may apply here. Although intravenous administration affords more immediate, potent, and economic effects as compared with oral and intranasal administration, it carries a higher probability of long-term adverse consequences such as abscesses, thrombosed veins, hepatitis, and overdose. Since most of these relatively youthful multiple drug abusers had not yet experienced the physical liabilities that accompany extensive intravenous drug use, their method preferences may have been influenced primarily by the short-term advantages of injecting drugs.

Internal evidence from the individual preference matrices suggested that method of administration may interact with the type of drug to determine preference. One common response pattern showed, for example, that intravenous opiates were preferred to intravenous cocaine, but intranasal cocaine was preferred to intranasal opiates. Since method of administration seems to have interactive effects as well as main effects on preference, research concerning drug use and preference should include distinctions among the different methods of administration using more comprehensive arrays of drug types and methods than have been investigated heretofore.

The independence of drug preference and frequency of use suggests several implications for the treatment of drug abusers. In cases in which the most preferred and most used drugs differ, knowledge of the patient's drug preferences could be useful in formulating an optimal course of rehabilitation. Although most

heroin users prefer heroin to all other drugs, for example, the addiction to heroin by a minority may obscure equally severe problems involving more preferred nonopiate drugs. Rehabilitative therapies that concentrate on the most frequently used drug may be less than successful if the existence of the preferred drug is not discovered and given a commensurate degree of clinical attention. Accurate measures of the applicant's drug preferences, administered on a routine basis, would be useful in diagnosing these otherwise latent drugs of abuse.

The special case of opiate users who prefer nonopiate drugs such as cocaine and amphetamine is particularly problematic. These individuals may have become addicted to heroin in conjunction with the use of their preferred drug. Some of them are misidentified as confirmed heroin addicts and become enrolled in programs in which they are maintained for indefinite periods on methadone, a highly addictive synthetic opiate, when detoxification from heroin and treatment primarily for dependence on the preferred drug might be a more effective therapy. The drug preference inventory could be used to identify prospective methadone patients whose preferred drugs are nonopiates.

Many rehabilitative programs rely on techniques intended to modify personality and other intrapsychic characteristics that are believed to underlie compulsive drug use. The rationale for these practices may be invalid to the extent that it depends on the assumption of a direct relationship between personality characteristics and drug use. These psychotherapeutic techniques also risk the danger of emphasizing the drug most frequently used

to the point of missing the significance of secondary drugs of abuse that may be more highly preferred. If a causal chain exists from personality characteristics through drug preference to drug use, then therapy might more effectively curtail compulsive drug use by attempting to change drug preferences rather than characteristics that are more remotely implicated in drug abuse.

References

- Baer, D. J., & Corroda, J. Age of initial drug use and subsequent preference for other drugs. *Psychological Reports*, 1973, 32, 936.
- Crain, W. C., Ertel, D., & Gorman, B. S. Personality correlates of drug preference among college undergraduates. *International Journal of the Addictions*, 1975, 10, 849-856.
- Edwards, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.
- Gossop, M. R., & Connell, P. H. Attitudes of oral and intravenous multiple drug users toward drugs of abuse. *International Journal of the Addictions*, 1975, 10, 453-466.
- Henriques, E., Arsenian, J., Cutter, H., & Samara-weera, A. B. Personality characteristics and drug of choice. *International Journal of the Addictions*, 1972, 7, 73-76.
- Penk, W. E., & Robinowitz, R. Personality differences of volunteer and nonvolunteer heroin and nonheroin drug users. *Journal of Abnormal Psychology*, 1976, 85, 91-100.
- Ungerer, J. C., Harford, R. J., Brown, F. L., Jr., & Kleber, H. D. Sex guilt and preferences for illegal drugs among drug abusers. *Journal of Clinical Psychology*, 1976, 32, 891-895.
- Zuckerman, M., Bone, R., Neary, R., Mangelsdorff, D., & Brustman, D. What is the sensation seeker? Personality trait and experience correlates of the Sensation-Seeking Scales. *Journal of Consulting and Clinical Psychology*, 1972, 39, 308-321.

Received May 27, 1977 ■

Neuropsychological Stability in Multiple Sclerosis

Robert J. Ivnik
Mayo Clinic and Mayo Foundation
Rochester, Minnesota

The neuropsychological performances of 14 patients who had multiple sclerosis (MS) and who received repeated testings spaced over time by at least 1 year were compared with identical evaluations of 14 patients who had neurological involvement but not MS. Subjects in each group were individually matched on chronological age at first testing, length of test-retest interval, sex, and years of formal education. Performance decrements attributable to the demyelination process of MS were primarily manifested on tasks requiring motor proficiency or complex sensory discriminations. Tests of higher order cognitive functions (e.g., abstractions, speech perception) were less adversely affected, except for measures having significant motor components. Preliminary Minnesota Multiphasic Personality Inventory data are also presented. The results indicate relative preservation or only mild deterioration for most intellectual abilities despite worsened motor-sensory functioning.

In recent years, the psychological test performances of patients with multiple sclerosis (MS) have been experimentally differentiated from the performances of either patients with neurological involvement but not MS or normal control patients. Matthews, Cleeland, and Hopper (1970) compared the test scores of 30 patients who had MS with those of a neurological group that excluded MS subjects on an extensive battery of neuropsychological tests and found significantly poorer functioning for the patients with MS on tests that demanded motor skill, speed, and coordination. Reitan, Reed, and Dyken (1971) reported similar findings when patients who had MS were compared with a neurologically normal control group. Reitan et al. also noted less striking but statistically significant defi-

cits on tests of verbal and auditory perception, measures of incidental recall for geometric figures and spatial relations, and tests of sensory-perceptual alertness. General information, verbal communication, and comprehension measures were least affected by MS. In each study, relatively good preservation of reasoning and logical-analytic skills was apparent. The validity of the studies by Matthews et al. and Reitan et al. is supported by Goldstein and Shelly's (1974) successful partial replication of each. Finally, Beatty and Gange (1977) provided correlative evidence suggesting that memory functions also may suffer the effects of demyelination, but this hypothesis at present is tentative and requires additional experimental verification.

MS is a disease manifested by exacerbations and remissions of clinical symptoms associated with a slowly progressive deterioration of general functioning. Even though neuropsychological studies have defined distinguishing behavioral-performance features of the disease, the relationship between the extent of neuropsychological impairment and the clinical progression of the disease has not been described. To approach this question, Ivnik (1978) compared the neuropsychological performances of three groups of patients:

This research was completed while the author was a postdoctoral fellow in clinical neuropsychology at the Neuropsychology Laboratory, University of Wisconsin.

The author is particularly appreciative of the assistance provided by J. Preston Harley and Charles G. Matthews during the course of manuscript preparation.

Requests for reprints should be sent to Robert J. Ivnik, Department of Psychiatry and Psychology, Mayo Clinic, Rochester, Minnesota 55901.

Table 1
*Neurological Diagnoses of 14 Non-MS
 Neurological Control Patients*

Patient	Diagnosis
1	Left cerebral atrophy; history of alcohol abuse
2	Major motor seizures secondary to subarachnoid hemorrhage
3	Posttraumatic syndrome
4	S/P encephalitis
5	Somatosensory-evoked seizures of unknown etiology
6	S/P neurosurgical repair of depressed skull fracture
7	S/P neurosurgical repair of epidural hematoma
8	S/P skull fracture with cerebral contusion
9	Major motor seizures of unknown etiology
10	Major motor seizures of unknown etiology
11	Parkinson's disease
12	Mixed seizure disorder of unknown etiology
13	Partial-complex seizure of unknown etiology
14	Mild mental retardation; radiculopathy; S/P myasthenic syndrome

Note. MS = multiple sclerosis; S/P = status/post.

those whose duration of MS was 1–5 years, 6–10 years, or greater than 10 years. Neuropsychological tests generated surprisingly few significant findings. Ivnik observed that the individual rate at which MS progresses is so varied that large-group comparative statistics may be an inappropriate experimental design for examining the disease process. To more rigorously examine the stability of neuropsychological performances over time for patients with MS, the research reported herein contained repeated examinations of the subject population using each patient as his or her own control. Patients with MS who were seen for extensive neuropsychological examinations on two occasions were compared with a population of patients with non-MS neurological involvement who also received repeated evaluations.

Method

Subjects

The testing protocol of every patient with MS who had received repeated examinations in the

Neuropsychology Laboratory, University of Wisconsin, was reviewed. Only patients with definite diagnoses of MS were considered; patients described as "possible MS," "probable MS," "MS suspect," or in any other equivocal manner were excluded. The referring neurologist's judgment that the diagnosis was certain determined group inclusion. If a patient's diagnosis was equivocal at the time of initial testing but definite when reevaluated, he or she was included for further consideration. A minimal test-retest interval of 1 year was required. Persons not receiving complete neuropsychological evaluations were excluded unless the missing data reflected a patient's physical disability due to the neurological illness. For example, if a patient's manipulative skills were so deteriorated that he or she was unable to perform the test of fine motor coordination, the patient was included in the study and the worst possible score was assigned for the missing test. If, however, data were missing for reasons unrelated to the ability being assessed (e.g., a patient did not receive the Category Test of nonverbal abstraction abilities because of poor visual acuity), no theoretic data were supplied. There were few instances in which data were supplied, because one criterion for selection in the study was having completed most of the neuropsychological measures.

Determination of the non-MS neurological control group required review of all neuropsychological patient protocols on file. Persons with incomplete test protocols, only one neuropsychological examination, or primary diagnoses of nonneurological conditions were excluded from further consideration. Each patient with MS was directly matched to a non-MS neurological control patient with regard to sex, chronological age when first tested, education, and number of months between testings.

The decision to use only non-MS neurological patients as controls was made to distinguish the stability of MS from that of other neurological disorders. From a clinical viewpoint, a neurological control group comprised of diseases that are most frequently included in the differential diagnoses with MS (e.g., spinal cord lesions) would have been optimal, but practical concerns made this impossible. The final MS and neurological control groups included 14 patients each. Neurological diagnoses of control patients are shown in Table 1.

Neuropsychological Tests

The following measures served as dependent variables in this study.

Wechsler Adult Intelligence Scale (WAIS). Verbal, Performance, and Full Scale IQs (Wechsler, 1955) were compared for each experimental group. Individual WAIS subtest performances also were analyzed.

Wide Range Achievement Test (WRAT). Word recognition, spelling, and arithmetic abilities were tested and scored for grade level equivalencies (Jastak & Jastak, 1965).

Category Test. A test of abstraction-concept formation ability using 208 visual stimuli was presented on a screen (Halstead, 1947). The number of errors was recorded.

Speech Perception Test. This is an auditory discrimination task in which the subject underlines on an answer sheet one of four nonsense syllables that most closely corresponds to the stimulus presented via a tape recorder (Halstead, 1947). The test was scored for number of errors.

Tactual Performance Test (TPT). A 10-block form board was given under a blindfolded condition. Average time (in minutes per block) for three consecutive trials (preferred hand, nonpreferred hand, both hands) was recorded, as was the total time for the three trials. The number of blocks that the subject could recall for shape (memory component) and correct location (location component) on a posttest drawing of the board and blocks was also noted (Halstead, 1947).

Trail Making Test. A paper-and-pencil test was given in which on Part A a subject connects as rapidly as possible 25 numbered circles distributed on a sheet of paper (Armitage, 1946). On Part B, half of the circles are numbered and half are lettered, and the subject connects the circles by alternating between these two sequences. The number of seconds to complete each part was the dependent measure.

Seashore Rhythm Test. A test of the subject's ability to make same-different discriminations between 30 pairs of rhythmic patterns was presented with a tape recorder (Seashore, Lewis, & Sætevit, 1960). The number of correct responses was recorded.

Imperception Test. Standardized examination was given for eliciting tactile, auditory, and visual imperception or suppression (or both) errors (Reitan, 1959b).

Finger Agnosia Test. The patient was required to identify by touch alone fingers of his or her hand after they were lightly touched with a pencil point; 20 trials were given on each hand, and the number of errors was recorded (Reitan, 1959b).

Fingertip Number Writing Test. The patient was blindfolded and was required to identify numbers written on his or her fingertips; 20 trials were given on each hand, with the number of errors being noted (Reitan, 1959b).

Tactile Form Discrimination Test. The subject identified by touch alone one of four plastic shapes placed in his or her hand, indicating his or her response by pointing to a display panel. Time and error scores were recorded for each hand (Reitan, 1959b).

Sandpaper Roughness Discrimination Test. The patient was blindfolded and was presented with four wooden blocks, each of which was covered with a different roughness of sandpaper. The subject was required to arrange (as quickly as possible) the sandpaper blocks in order of increasing roughness. Time and error scores were computed for each hand. 1959b).

Grooved Pegboard Test. A manipulative dexter-

Table 2

Means and Standard Deviations of Matching Variables for MS Group and Non-MS Neurological Control Group

Variable	MS	Control
Chronological age		
at first testing	38.0 \pm 9.28	37.0 \pm 9.32
Education (years)	12.5 \pm 2.65	12.1 \pm 2.80
No. months		
between testings	37.0 \pm 28.6	34.4 \pm 29.7

Note. MS = multiple sclerosis. Sex was directly matched. There were no significant differences.

ity task using a pegboard containing 25 holes with randomly positioned slots (Lafayette Instrument Co., Model 4202) was given. Before they could be inserted, the pegs, with an edge along one side, needed to be rotated to match the hole. Time scores were recorded for each hand.

Maze Coordination Test. A modified 2706 A Maze (Lafayette Instrument Co.) was placed on a stand in a vertical position at the subject's midline. The subject was required to go through the maze with an electric stylus, trying not to touch the sides. The stylus was attached to a time clock and counter. Cumulative time of contact with the maze and cumulative error scores were recorded for each hand.

Static Steadiness Test. The subject inserted an electric stylus into the holes of a conventional hole-type steadiness test (Lafayette Instrument Co., Model 4605 C). The subject was asked to keep the stylus in each hole for 15 sec. Cumulative time and counter scores were recorded for each hand.

Finger Tapping Speed Test. The subject was required to tap (as fast as possible) his or her index finger on a counter apparatus. The mean of five 10-sec trials was recorded for each hand (Halstead, 1947).

Wisconsin Impairment Index. A composite score was computed for each subject, ranging from 0 to 1.0, based on the frequency with which the subject exceeded specified cutoff points on 10 tests routinely used in the neuropsychological examination. Six of 10 Impairment Index measures were represented by tests from Halstead's battery. The other four measures were (a) time score on Part A plus Part B of the Trail Making Test exceeding the cutoff point suggested by Reitan (1958); (b) two or more definite dysphasic symptoms on a modified and extended version of the Halstead-Wepman Aphasia Screening Test (Halstead & Wepman, 1949); (c) distorted reproductions of square, triangle, or Greek cross figures on the Aphasia Screening Test; and (d) occurrence of two or more errors on one body side in tests of tactile finger identification and fingertip number writing perception, or the presence of two or more lateralized errors on sensory imperception testing in tactile, auditory, or visual modalities.

Table 3
Raw Score Descriptive Statistics on All Dependent Measures

Measure	MS testing				Control			
	First		Second		First		Second	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
WAIS								
Full	103.4	10.8	99.7	13.3	98.4	15.7	100.4	15.0
Verbal	107.7	10.8	104.1	13.4	99.4	18.6	100.9	17.4
Performance	97.7	12.8	94.0	14.6	96.6	15.3	99.7	13.1
WAIS subtests								
Information	11.5	2.5	11.0	2.4	9.9	2.6	10.5	2.6
Comprehension	12.0	2.8	11.5	3.1	10.5	3.4	10.1	2.8
Arithmetic	11.4	3.2	10.8	3.7	9.7	4.0	10.2	4.3
Similarities	11.1	2.1	10.6	2.7	10.5	2.8	10.6	2.9
Digit Span	10.5	1.5	9.2	2.5	9.4	3.2	8.6	3.6
Vocabulary	11.0	2.6	10.3	2.6	9.7	3.1	10.1	3.2
Digit Symbol	7.5	2.4	5.9	2.6	6.8	2.2	7.3	2.0
Picture Completion	10.0	1.8	9.9	2.5	9.4	2.5	10.0	2.3
Block Design	9.7	1.6	8.3	3.1	9.8	3.1	9.6	3.0
Picture Arrangement	8.8	2.7	8.1	2.3	9.2	2.4	8.0	2.4
Object Assembly	7.6	3.1	7.7	3.3	8.7	2.9	10.0	2.0
WRAT (grade level)								
Reading	11.8	4.2	11.3	3.7	9.5	4.5	9.6	4.3
Spelling	8.7	3.0	8.6	2.8	7.4	2.9	7.2	3.1
Arithmetic	7.8	2.9	7.6	3.4	6.0	2.9	5.3	1.8
Impairment Index	.55	.27	.68	.23	.64	.23	.53	.25
Category Test (errors)	52.3	24.1	51.1	21.8	58.4	20.9	49.4	27.5
Seashore Rhythm Test (number correct)	24.1	3.1	23.4	4.2	22.2	4.2	22.5	4.1
Speech Perception (errors)	5.8	3.4	6.1	2.3	10.1	7.0	7.9	7.2
Trails A (seconds)	40.1	11.8	55.6	29.5	56.0	44.7	51.9	22.1
Trails B (seconds)	94.9	32.0	178.2	126.7	135.4	96.9	137.1	9.1
TPT								
Dominant (min/block)	2.3	2.4	3.0	3.2	1.9	1.5	1.3	1.2
Nondominant (min/block)	1.9	2.4	3.3	3.2	1.9	2.5	1.5	1.7
Both (min/block)	2.1	3.0	1.6	2.5	1.8	2.6	.8	1.1
Total Time (minutes)	39.1	24.0	47.6	28.6	45.9	35.5	31.4	29.0
Memory	5.6	2.3	5.9	1.6	5.7	2.9	6.3	2.0
Location	2.8	2.1	2.1	1.4	2.1	2.4	3.3	2.5
Imperception Testing (errors)								
Tactile	.36	.84	.64	1.50	.29	.61	.86	1.88
Auditory	.14	.53	.86	1.75	.64	1.22	.79	2.12
Visual	.14	.53	.86	1.61	.07	.27	.29	.61
Total	.64	1.15	2.36	3.32	1.00	1.30	1.93	3.87
Finger Agnosia (errors)								
Right	1.3	1.8	1.1	1.6	1.6	2.9	1.2	2.0
Left	1.5	1.7	1.4	2.2	1.2	1.6	1.1	2.0
Total	2.79	2.89	2.42	3.34	2.79	3.89	2.29	3.75

Table 3 (continued)

Measure	MS testing				Control			
	First		Second		First		Second	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Fingertip Number Writing (errors)								
Right	2.9	4.1	5.6	6.2	2.9	2.8	2.2	2.8
Left	2.6	4.0	4.1	5.6	2.0	3.9	1.1	2.0
Total	5.5	8.1	9.7	10.5	4.9	6.5	3.4	4.6
Sandpaper (sec)								
Right	33.7	8.2	52.7	28.0	47.9	25.0	38.6	17.0
Left	43.3	20.1	55.4	31.8	39.4	20.0	35.5	13.7
Total	78.6	30.1	109.1	70.0	90.6	51.1	71.1	30.7
Sandpaper (errors)								
Right	.57	0.94	1.00	1.30	1.14	1.87	.57	1.22
Left	.71	1.68	1.14	1.92	1.00	1.88	.71	1.27
Total	1.28	2.30	2.14	3.58	2.14	3.63	.71	1.27
Tactile Forms (sec)								
Right	28.8	23.6	38.1	25.7	25.9	7.3	20.2	6.1
Left	31.1	23.1	40.6	28.8	21.6	6.4	17.4	5.2
Total	60.1	44.6	78.7	53.2	47.5	13.0	40.6	13.8
Tactile Forms (errors)								
Right	.57	.75	.79	1.42	.29	.82	.07	.27
Left	.86	2.07	1.86	2.63	.07	.27	.00	.00
Total	1.43	2.47	2.64	3.30	.36	1.08	.64	2.13
Finger Tapping (count)								
Dominant	43.4	9.0	37.2	7.6	39.7	10.1	44.1	10.2
Nondominant	36.2	7.7	29.8	10.9	35.0	9.0	38.9	5.6
Dynamometer (kg)								
Dominant	44.7	9.0	39.4	9.8	36.7	11.4	41.8	15.6
Nondominant	38.1	14.1	33.8	15.1	35.9	10.6	37.9	14.2
Pegboard (sec/peg)								
Dominant	4.3	1.8	5.6	2.5	3.8	.9	3.6	.9
Nondominant	5.9	3.9	25.6	40.3	4.2	1.2	4.0	.9
Mazes (sec)								
Dominant	4.31	4.24	6.21	6.52	4.39	6.57	2.39	3.41
Nondominant	14.61	25.65	22.06	34.01	6.75	5.70	4.65	6.80
Mazes (count)								
Dominant	27.1	27.0	37.6	36.8	23.3	24.0	15.9	18.0
Nondominant	106.1	253.9	182.6	347.7	42.6	38.4	25.4	30.2
Static Steadiness (sec)								
Dominant	22.62	18.00	34.93	25.33	22.82	9.36	23.62	14.79
Nondominant	34.62	27.35	42.39	30.65	32.82	20.10	31.91	19.57
Static Steadiness (count)								
Dominant	109.1	75.9	139.3	67.7	155.9	90.5	126.3	68.7
Nondominant	112.1	53.3	259.7	320.8	155.9	73.4	143.7	64.7

Note. WAIS = Wechsler Adult Intelligence Scale; WRAT = Wide Range Achievement Test; TPT = Tactual Performance Test.

Minnesota Multiphasic Personality Inventory (MMPI). Ten patients in each group also completed the MMPI at each testing. Their profiles were compared for each scale as a related brief study.

Data Analysis

Normative data for many of the above-described tests have been provided by Kiernan and Matthews (1976). The availability of these norms permits the transformation of raw scores on individual tests into *T* scores ($M = 50$, $SD = 10$), thereby providing a mechanism for comparing the experimental subjects' performances against "normal expectation." *T*-score conversions also allow for direct comparison of performance levels in various ability domains because raw score measurement units (e.g., kilograms for grip strength, IQ for intelligence) are transformed to a common scale. Further, these conversions provide the opportunity for a graphic display of the test results. A disadvantage of Kiernan and Matthews' results is that normative data on sensory examinations were not included. The dependent

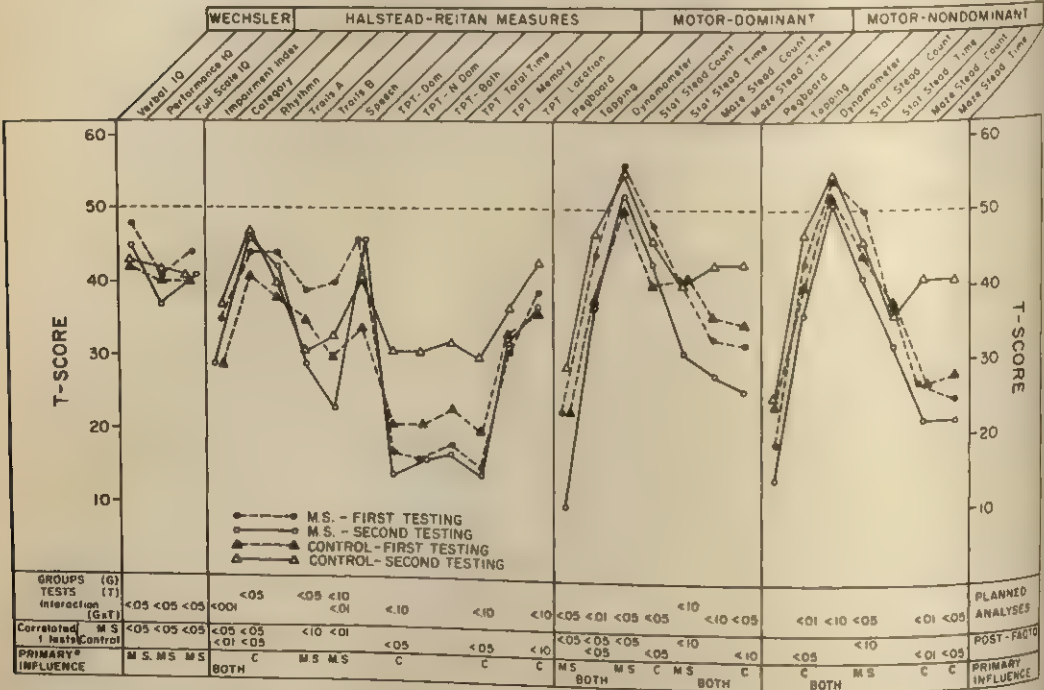
variables to which the *T*-score conversions apply include Verbal, Performance, and Full Scale IQs from the WAIS, Wisconsin Impairment Index, Category Test, Seashore Rhythm Test, Trail Making Test, Speech Perception Test, Tactual Performance Test, and all tests of motor-steadiness proficiency.

Analyses of variance (ANOVA) (2×2) were computed with repeated measures on the second variable (i.e., testings). On measures for which the ANOVA identified a trend ($.05 < p < .10$) or a significant difference ($p < .05$), post facto *t* tests for correlated means were computed for each group across testings. The comparisons were of specific interest when the ANOVA identified a significant interaction, because the *t*-test analyses provided information as to which patient groupings (i.e., either one or both) primarily influenced the interaction.

Results

Both patient groups were comparable in chronological age at first testing, length of

M.S. AND CONTROL GROUP PERFORMANCES AT FIRST AND SECOND TESTING ON DEPENDENT MEASURES WITH T-SCORE CONVERSIONS



*Reflects the author's interpretation as to whether the change in group mean performance of the M.S., control, or both groups acting together was primarily responsible for the identified significance levels

Figure 1. Multiple sclerosis (M.S.) and control group performances at first and second testings on dependent measures, with *T*-score conversions. [Results of a 2×2 analysis of variance (ANOVA) on each variable are shown at bottom, and only those analyses that showed $p < .10$ significance are listed. Post facto correlated *t* tests comparisons across testings also are given on those measures in which ANOVAs identified a significant test or interaction effect.]

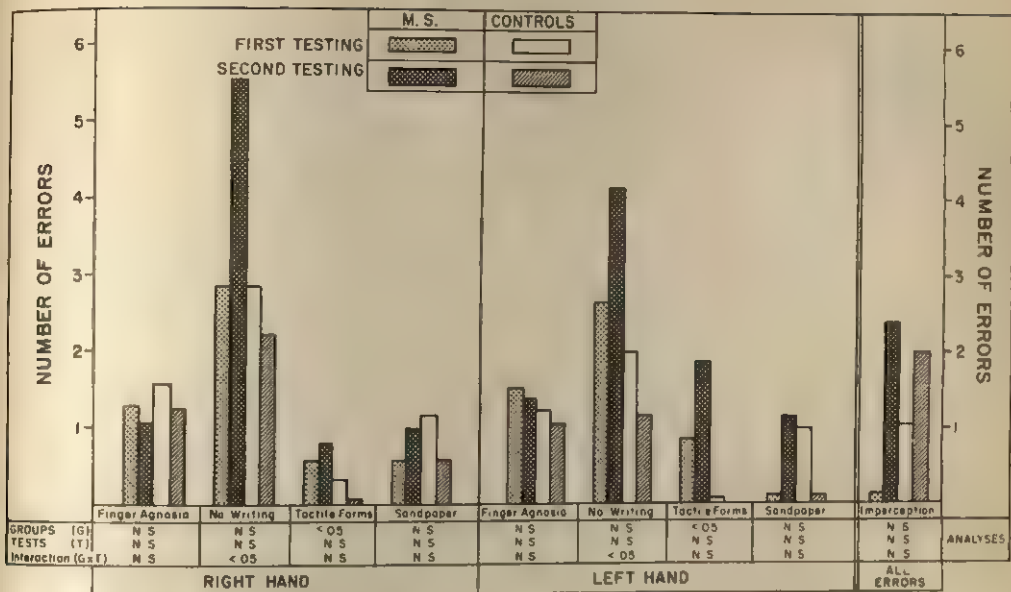


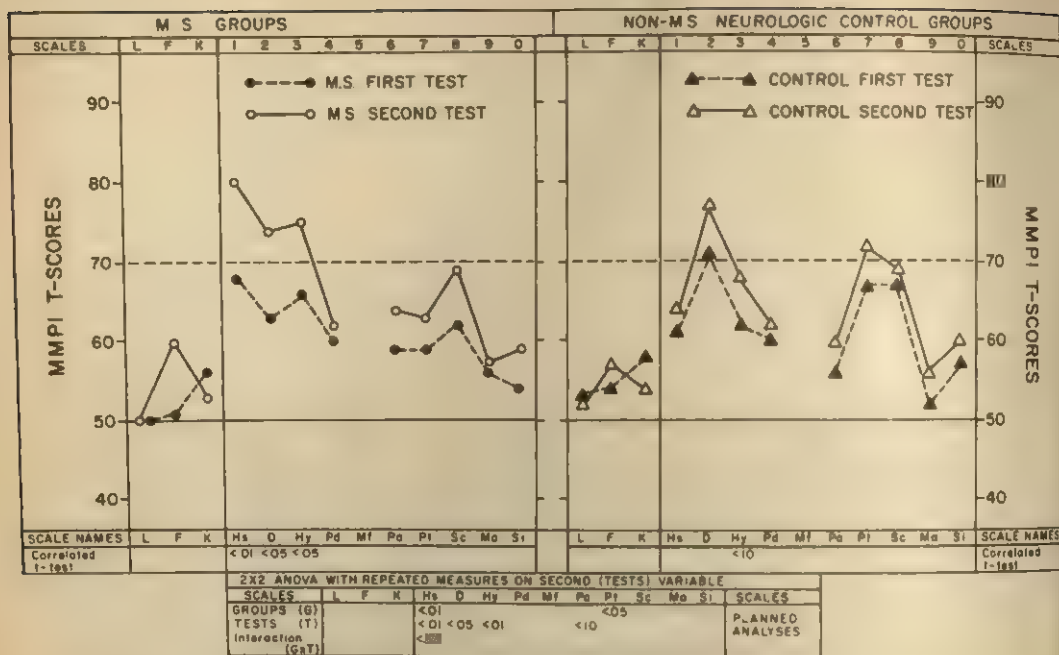
Figure 2. Multiple sclerosis and control group performances on sensory-perceptual examinations.

test-retest interval, sex, and years of formal education (Table 2).

Evaluation of descriptive raw data of each subject group at each testing before any unit of measurement (i.e., *T*-score) conversions or statistical analyses revealed extremely impaired mean values and large standard deviations on several of the measures, particularly motor performances (Table 3). These extreme values most frequently represented spurious exaggeration of group mean values partly because one or more patients could not complete the test owing to neurological dysfunction, and the patient therefore was assigned the worst score possible for that measure. *T*-score conversions served to attenuate the undue influence of such artificially extreme values.

Analyses of conversion of raw data to *T*-scores revealed significant differences between the groups (see Figure 1). On measures in which significant differences were found, the MS patient group's performances were either unchanged or worsened over time, but they never significantly improved. In contrast, the performance of the non-MS neurological controls was either consistently unchanged or improved, but it never significantly worsened.

The number of imperception errors in each sensory modality (i.e., tactile, auditory, visual, and total number of errors across modalities) failed to distinguish MS patients from controls (see Figure 2). The Sandpaper Roughness Discrimination and Finger Agnosia examination results also failed to show significant group differences when the number of errors was used as the dependent measure; however, when the time taken to complete the Sandpaper Roughness Discrimination Test was examined, a significant interaction effect for the right hand ($p < .01$) was evident, and left-handed performances approached significance ($.05 < p < .10$). The performance of the MS group on this measure noticeably deteriorated bilaterally, whereas the control group clearly improved. There was a significant group effect ($p < .05$) in the sensitivity of number of errors on the Tactile Forms Discrimination Test to MS, but no testing or interaction effects were apparent (see Figure 2). The time component of the Tactile Forms Discrimination Test identified a significant right-handed interaction effect ($p < .05$) and a significant left-handed group effect ($p < .01$). Right-handed performances analyzed across groups and left-handed performances on interaction



analyses approached significance ($.05 < p < .10$) on this measure. These results suggest fairly good sensitivity to MS deterioration for the error and time components of the Tactile Forms Discrimination Test. Finally, the Fingertip Number Writing Test generated consistent interaction effects for number of errors on both the right and left hands, along with the total number of errors for this test. This interaction was chiefly caused by the worsened performance of the MS group and by the mild improvement demonstrated by the controls.

Because WAIS Verbal, Performance, and Full Scale IQs yielded significant interaction effects, analyses were also computed on the individual WAIS subtests (Table 3). Among the verbal subtests, significant interactions without significant group or testing effects were obtained on the Information ($p < .01$) and Vocabulary ($p < .05$) subtests; Digit Span showed a testing effect ($p < .01$), but no group or interaction effects were apparent. Among the performance subtests, Digit Symbol also yielded a significant interaction effect ($p < .01$) without group or testing differences; Picture Arrangement attained a sig-

nificant ($p < .05$) testing effect only; and Block Design approached significance ($.05 < p < .10$) on the analyses of tests across groups.

The WRAT failed to show significant differences for any academic skill.

The test-retest MMPI profiles for each experimental group revealed some statistically significant scale differences (see Figure 3).

Discussion

This study introduces a number of issues regarding the changes over time for cognitive and motor-sensory function in MS. MS deterioration continues to be most apparent on measures that demand coordinated motor skills and on cognitive ability tests having a significant motor component. The Grooved Pegboard Test (fine motor coordination and manual dexterity) was the most poorly performed test in the MS group on initial testing and showed the greatest decrement over time. Kinetic steadiness (Maze Coordination Test) was also definitely impaired, whereas finger tapping speed and static steadiness generally were borderline or only mildly impaired (i.e.,

less than 2 SDs below normal expectation). Grip strength decreased slightly among MS patients but remained average. Among tests of "higher order" abilities, only the Trail Making Test, which is motorically dependent, showed significantly worsened scores for MS patients. The TPT was poorly executed by both the MS and control subjects; however, control subjects showed improvement, whereas the performances of the MS patients were essentially unchanged. The worsening on motor proficiency measures for patients with MS was consistent with previous research (Goldstein & Shelly, 1974; Matthews et al., 1970; Reitan et al., 1971).

An unexpected finding was that the motor deterioration for MS patients appeared to be more evident on dominant than nondominant hand measures. This suggestion evolves from the post facto *t* test analyses in which dominant hand test-retest comparisons on the pegboard, finger tapping, and static steadiness measures either approached or achieved significance, but nondominant hand performances on these same tests did not. On several of these motor tests (e.g., finger tapping and steadiness measures), the control subjects showed bilateral improvement. The clinical or neuropathological (or both) significance of this observation is uncertain, but it raises the possibility that deterioration associated with MS is more apparent, and subsequent referral for neuropsychological reevaluation is therefore pursued, as the efficiency of dominant hand performance is affected—suggesting a possible sampling bias among the MS patients.

Several of the tests of sensory discrimination abilities also appear to be sensitive to MS. The number of errors on the Fingertip Number Writing Test increased dramatically in the follow-up MS data, as did the time scores on the Sandpaper Roughness Discrimination Test. The stereognostic abilities of the MS patients were significantly worse than those of control subjects, both in initial and in follow-up testing, but no statistically significant deterioration was identified. In considering these data, it must be appreciated that an unknown degree of the motor-sensory deterioration seen in MS may reflect plaque

formation below the level of the cerebral hemispheres.

In addition to the worsened performances over time on motor-dependent measures and several sensory-perceptual tests, the MS group consistently showed either unchanged or worsened scores on all other tests of adaptive abilities. This observation was in sharp contrast to the opposite tendency for non-MS neurological control subjects to improve on many of these same tasks. Improvement by the control group may indicate improvement of adaptive skills, a "practice effect," or both. The fact that the mean test-retest interval was almost 3 years suggests that any practice effect would most likely be minimal. The probability that the test-retest differences represent a true improvement in neuropsychological function for the control group is supported by the fact that several of the neurological disorders of the control group involved neurosurgical procedures from which recovery of function over time could reasonably be anticipated.

Neuropsychological decrements over time were also evident on several tests of higher order cognitive skills. Verbal, Performance, and Full Scale IQs on the WAIS achieved statistical significance, as did the Wisconsin Impairment Index. The WAIS measures were primarily influenced by worsening of MS, whereas the Impairment Index reflected not only the exacerbation of adaptive ability deficits in the MS group but also improved performances by the non-MS neurological controls. The worsened Impairment Index ratings for MS patients seem to be largely attributable to the changes evident on the Trail Making Test and the TPT, each of which is heavily motorically dependent. The Seashore Rhythm and the Speech Perception tests yielded no significant findings, but the Category Test results demonstrated a mild improvement over time that was most prominent for the control subjects.

It is somewhat surprising that WAIS variables should prove to be sensitive to the effects of demyelination when many other measures that are commonly considered to be more sensitive indices of brain dysfunction (Reitan, 1959a) yielded negative or equivocal

results. These unanticipated psychometric findings are further highlighted by the observation that the Information, Vocabulary, and Digit Symbol WAIS subtests were the measures that proved to be most sensitive to MS deterioration. The worsening of Digit Symbol scores in the MS sample may be related to this test's dependency on a visual-motor response, but worse scores on the Information and Vocabulary subtests cannot be explained on a visual-motor basis. Furthermore, these two subtests are commonly considered to be relatively insensitive to neurological disorders (in the absence of aphasia). Several WAIS subtests that are more often sensitive to neurological dysfunction failed to yield significant decrements for the MS group. The Digit Symbol, Picture Arrangement, and Block Design subtests either approached or achieved statistical significance on the analyses of tests across groups, with slightly worsened functioning evident on these measures regardless of neurological diagnosis. Even though it is of interest to note that many of the individual WAIS subtests, in addition to the summary IQ scores, achieved statistical significance, the clinical import of these findings may be minimal, because the absolute mean values on these measures were not strikingly changed in either diagnostic group. Although other tests of complex cognitive functions yielded statistically significant results, the actual level of impairment was not strikingly changed when the various neurological patient performances were compared against expectations based on normative data (Figure 1).

One ability domain that Beatty and Gange (1977) suggested as being particularly sensitive to demyelination is memory. Several different tests requiring attention-memory skills in this study found no support for significant memory impairment, but it must be remembered that the tests used here were brief clinical assessment procedures and were not the more rigorous techniques used in the experimental learning-memory literature.

The MMPI data are interesting for their implications for future research, but the current sample size was too small to permit any conclusions. Group profiles suggest an in-

creased concern with physical integrity that may be more frequent in MS. This concern probably reflects the actual disease process of MS. Neurological dysfunction in general appears to be associated with worsening depression, increased sensitivity to perceived or real criticism, and increased denial. Finally, there is the suggestion that MS patients may be less anxious than non-MS neurological controls. Future research into the MMPI correlates of MS will hopefully correlate group or individual (or both) MMPI profiles with clinical-behavioral status at the time of testing.

This study offers some encouragement to the MS patient and to professionals active in the patient's therapeutic and social-vocational planning. The generality of any conclusion drawn from this research is limited to those MS patients who remain capable of taking neuropsychological examinations. Nevertheless, the knowledge that one can reasonably expect to retain much of the premorbid intellect over an extended time interval may serve to keep the realistic anticipation of worsened motor skills and sensory-perceptual abilities in a broader and less catastrophic personal-social-vocational perspective.

References

- Armitage, S. G. An analysis of certain psychological tests used for the evaluation of brain injury. *Psychological Monographs*, 1946, 60, 1-48.
- Beatty, P. A., & Gange, J. J. Neuropsychological aspects of multiple sclerosis. *Journal of Nervous and Mental Disease*, 1977, 164, 42-50.
- Goldstein, G., & Shelly, C. H. Neuropsychological diagnosis of multiple sclerosis in a neuropsychiatric setting. *Journal of Nervous and Mental Disease*, 1974, 158, 280-290.
- Halstead, W. C. *Brain and intelligence: A quantitative study of the frontal lobes*. Chicago: University of Chicago Press, 1947.
- Halstead, W. C., & Wepman, J. M. The Halstead-Wepman Aphasia Screening Test. *Journal of Speech and Hearing Disorders*, 1949, 14, 9-15.
- Ivnik, R. J. Neuropsychological test performance as a function of the duration of MS-related symptomatology. *Journal of Clinical Psychiatry*, 1978, 39, 304-307; 311-312.
- Jastak, J. F., & Jastak, S. R. *The Wide Range Achievement Test: Manual of instructions*. Wilmington, Del.: Guidance Associates of Delaware, 1965.
- Kiernan, R. J., & Matthews, C. G. Impairment index

- versus *T*-score averaging in neuropsychological assessment. *Journal of Consulting and Clinical Psychology*, 1976, 44, 951-957.
- Matthews, C. G., Cleeland, C. S., & Hopper, C. L. Neuropsychological patterns in multiple sclerosis. *Diseases of the Nervous System*, 1970, 31, 161-170.
- Reitan, R. M. Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills*, 1958, 8, 271-276.
- Reitan, R. M. The comparative effects of brain damage on the Halstead Impairment Index and the Wechsler-Bellevue Scale. *Journal of Clinical Psychology*, 1959, 15, 281-285. (a)
- Reitan, R. M. *Manual for administration of screening test for aphasia and related perceptual disorders*. Indianapolis: Indiana University Medical Center, Neuropsychology Laboratory, 1959. (b)
- Reitan, R. M., Reed, J. C., & Dyken, M. L. Cognitive, psychomotor, and motor correlates of multiple sclerosis. *Journal of Nervous and Mental Disease*, 1971, 153, 218-224.
- Seashore, C. E., Lewis, D., & Saetveit, J. G. *Seashore measures of musical talents (manual)*. New York: Psychological Corporation, 1960.
- Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation, 1955.

Received June 10, 1977 ■

The Generalized Expectancy for Success Scale— A New Measure

Bobbi Fibel and W. Daniel Hale
University of Massachusetts

A new measure of generalized expectancy for success was assessed for its psychometric properties. Three samples of Caucasian, middle-class college students participated in the study. The first sample ($n = 100$; 59 females, 41 males) received a preliminary version of the Generalized Expectancy for Success Scale (GESS). Item analysis yielded 30 items that were substantially correlated with the total score but were not significantly related to social desirability. The second sample ($n = 104$; 63 females, 41 males) received the 30-item GESS twice at a 6-week interval. The third sample ($n = 103$; 69 females, 34 males) received the GESS, the Marlowe-Crowne Social Desirability Scale, the Internal-External Locus of Control Scale, the Self-Rating Depression Scale, the Depression Inventory, and the Hopelessness Scale. Results indicate that the GESS has acceptable test-retest reliability, high internal consistency, and minimal relationship with social desirability. Predicted relationships between high generalized expectancy for success, depressive symptomatology, and internality were supported. Factor analysis indicated that GESS scores are a function of one general factor. Further construct validation is reviewed, and implications for future use of the GESS are discussed.

One of the key concepts of Rotter's learning theory (Rotter, 1954; Rotter, Chance, & Phares, 1972) that has been the subject of considerable study in recent years is that of generalized expectancies. The two generalized expectancies that have received the most attention, and for which there are reliable and valid measures, are expectancies regarding internal-external control of reinforcements (Lefcourt, 1976; Phares, 1976; Rotter, 1966, 1975; Strickland, 1977) and interpersonal trust (Rotter, 1967). The purpose of the present ongoing investigation is to construct and validate a measure of a different generalized expectancy—the generalized expectancy for success. This construct can be de-

fined as the expectancy held by an individual that in most situations he/she will be able to attain desired goals. According to social-learning theory, an individual's behavior potential is a function of reinforcement value and expectancies that are determined by a person's reinforcement history for relevant situations. Therefore, when other factors are held constant, the behavior potential for an individual with a high expectancy for success should be greater than that of an individual with a low expectancy for success. Further, since situations vary in the extent to which a person's reinforcement history is relevant, expectancies for success may vary along a continuum from relatively specific to general, as a function of the degree of situational novelty or ambiguity. Numerous studies (Dickstein & Kephart, 1972; Feather, 1966; Feather & Saville, 1967; Rosenthal & Jacobson, 1966; Tyler, 1958) have demonstrated that individuals experimentally given a high expectancy for success on a certain task or set of tasks are indeed more likely to perform more successfully than those given a

This research was supported in part by funds available through National Institute of Health Grant HD 08952, principal investigator, Bonnie R. Strickland.

A portion of this article was presented at the meeting of the Eastern Psychological Association, New York, April 1976.

Requests for reprints should be sent to Bobbi Fibel, 255 West Elm Street, New Haven, Connecticut 06515.

low expectancy for success. Unlike these experimentally induced task-specific expectancies, however, situations seldom provide individuals with explicit expectancies for success. More commonly, individuals face relatively unfamiliar or ambiguous circumstances for which no highly specific expectancy has been provided or formulated. A person's behavior in such situations is still largely influenced by his/her expectancy for success, but this expectancy increasingly becomes a function of generalized expectancy as the degree of novelty or ambiguity increases. Just as the construction of a measure of internal-external locus of control allows researchers to move from situationally induced skill versus chance expectancies for control to internal-external control as an individual difference variable, the Generalized Expectancy for Success Scale (GESS) allows researchers to explore individual differences as a function of generalized expectancies for success.

A valid and reliable measure of a generalized expectancy for success can facilitate the study of factors in the development of such expectancies, situational characteristics that influence expectancies, and the impact of a generalized expectancy for success on a variety of goal-oriented behaviors and other theoretically related cognitive constructs. Thus, the GESS can potentially enhance prediction and clarify issues of theoretical importance. The development of such a scale depends not only on firm grounding in psychological theory but also on adherence to sound psychometric principles and extensive construct validation.

Method

Subjects

Three samples were obtained, each from large undergraduate psychology classes at a large, northeastern university. Students, predominantly middle-class Caucasians, were given the option to participate in studies of their own choosing for bonus academic points. A preliminary version of the GESS was administered to the first sample ($n=100$; 59 females, 41 males) during a class period. The second sample ($n=104$; 63 females, 41 males), solicited in the same manner, was group tested in the 4th and 10th weeks of the semester during the first 20 minutes of the class periods. In the third sample ($n=103$; 69 females, 34 males), subjects volunteered for one

of several small group-testing sessions based on the preliminary description given by the experimenters, one male and one female Caucasian graduate student.

Test Construction

Initially, an attempt was made to construct items that both sampled across situational domains (such as public, private, familial, interpersonal, and work related) and did not specify criteria for success. One hundred fifty items were constructed by the experimenters. The 150 items were then screened for face validity by three psychologists. One hundred four items were selected and subsequently administered to the first sample of 100 subjects. An item analysis yielded 30 items that were substantially correlated with total score ($r > .50$) but were not significantly related to social desirability ($p > .10$) as measured by the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960). These 30 items constitute the current version of the measure (see Appendix). All items begin with the same stem phrase: "In the future I expect that I will . . ." which is printed at the top of each page. Responses to items are in Likert format. Subjects are instructed to circle a number on a 5-point scale from 1 (highly improbable) to 5 (highly probable) for each item. Seventeen items are phrased in the positive or success direction and 13 in the negative or failure direction. Items are randomly ordered. The scale is scored additively and in the direction of success, such that a high total scale score indicates a high expectancy for success.

Procedure

The second sample of subjects ($n=104$) was run during fall 1975 and received only the GESS on two occasions for test-retest reliability purposes. The third sample of subjects, run in groups of 10-15 during spring 1975, received the GESS, the Marlowe-Crowne Social Desirability Scale, Rotter's Internal-External Locus of Control (I-E) Scale (Rotter, 1966), the Self-Rating Depression Scale (Zung, 1965), Beck's Depression Inventory (Beck, 1967), the Hopelessness Scale (Beck, Weissman, Lester, & Trexler, 1974), and a questionnaire assessing suicidal ideation (Crepeau, Note 1). Responses to the 30 GESS items were intercorrelated, and the resulting matrix was factored by the principal components method. Components were rotated to orthogonal simple structure by means of Kaiser's (1958) varimax method. Minimum eigenvalue for factor rotation was 1.50.

Results

The test-retest correlation coefficient of the GESS using scores taken at a 6-week interval from subjects in the second sample

Table 1
Correlations Between the GESS and Selected
Other Measures for the Present Study

Scale	Males		Females	
	<i>n</i>	<i>r</i>	<i>n</i>	<i>r</i>
Social Desirability	34	.15	69	.26*
Self-Rating Depression	26	-.58**	58	-.48**
Depression Inventory	25	-.61**	57	-.54**
Hopelessness	26	-.69**	59	-.31**
Locus of Control	32	-.10	67	-.27*

Note. GESS = Generalized Expectancy for Success Scale.

* $p < .05$.

** $p < .01$.

who were present for both administrations ($n = 74$; 46 females, 28 males) was .83 overall (.89 for males and .80 for females). Means and standard deviations on the GESS were not significantly different for this sample from those obtained as a function of group testing with additional measures, nor were differences in responding found as a function of sex. Consequently, data from the second and third samples were combined for analyses of psychometric properties ($n = 207$, 132 females, 75 males). The possible range of total scores is 30–150, with higher scores indicating a high expectancy for success. Actual total scores ranged from 65 to 143 for females and from 81 to 138 for males. The mean score for females was 112.32 (mode = 112, $Mdn = 113.14$) and for males, 112.15 (mode = 109, $Mdn = 112.88$). The respective standard deviations were 13.80 and 13.24.

Two measures of internal consistency were computed. The split-half reliability coefficient for odd versus even items, using the Spearman-Brown correction formula, was .90 for females and .91 for males. The correlation between the first 15 items and the last 15 items, again using the Spearman-Brown correction formula, was .82 for females and .83 for males. In view of the fact that all items have a single stem, the high internal consistency is not surprising. However, it should be noted that these reliability coefficients also occur across items that reflect a number of diverse areas.

Correlations with other measures were computed separately for males and females in the third sample. For both sexes it was found that scores on the GESS were correlated negatively and significantly with scores on the Zung Self-Rating Depression Scale, the Beck Depression Inventory, and the Beck Hopelessness Scale (see Table 1). Individuals with low expectancies for success were more likely to report depressive symptomatology and to report themselves as feeling hopeless about impending life events. Even though items with high social desirability bias were eliminated from the original item pool, a low but significant correlation between scores on the GESS and scores on the Marlowe-Crowne Social Desirability Scale was found for females ($r = .25$, $p < .02$) but not for males ($r = .15$, $p > .10$) in the third sample. Also, high GESS scores were significantly correlated with internality as measured by Rotter's I-E scale for females but not for males.

A factor analysis was also computed using the varimax rotation method, which yielded four factors. Variance accounted for by each factor was 63.9%, 13.4%, 12.7%, 10.1% for Factors 1–4, respectively.

Items loading on Factor 1 reflect an individual's sense of general efficacy (Items 4, 8, 9, 10, 12, 13, 15, 16, 21, and 22). For example, the two items with the highest loading within Factor 1 were Item 4, "be unable to accomplish my goals" (.56), and Item 21, "succeed at most things I try" (.55).

Factor 2 was composed of Items 14, 17, 24, 25, 26, 29, and 30. The content of these items primarily involves long-range career-oriented expectancies. Items with the highest loading within Factor 2 were Item 26, "attain the career goals I have set for myself" (.56), and Item 30, "achieve recognition in my profession" (.53).

Factor 3 contained items related to personal problem solving (3, 5, 6, 11, 19, 20, 23, 28). Item 5, "have a successful marital relationship" (.59), and Item 29, "be very successful working out my personal life" (.51), loaded highest among the items in the third factor.

Factor 4 consisted of Items 1, 2, 7, 18, and 27 with Item 1, "find that people don't seem to understand what I am trying to say" (.55),

Table 2

Summary of Correlations Between the GESS and Measures of Depressive Affect and Cognition from Other Studies

Scale	Males		Females		Total	
	<i>n</i>	<i>r</i>	<i>n</i>	<i>r</i>	<i>n</i>	<i>r</i>
Self-Rating Depression						
Strickland (Note 2)	50	-.55***	50	-.43***		
Crepeau (Note 1)	67	-.74**	107	-.62***		
Koerner (1977)					120	-.47***
With <i>SD</i> partialled out					120	-.44***
Multiple Affect Adjective Checklist						
Strickland (Note 2)						
Anxiety	50	-.33**	50	-.27*		
Depression	50	-.34**	50	-.20		
Hostility	50	-.03	50	-.04		
Crepeau (Note 1)						
Depression	67	-.66***	107	-.45***		
Koerner (1977)						
Anxiety					120	-.43***
Depression					120	-.37***
Hostility					120	-.41***
Suicidal Ideation						
Crepeau (Note 1)	67	-.33**	107	-.48***		

Note. GESS = Generalized Expectancy for Success Scale.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

and Item 18, "find that no matter how hard I try, things just don't turn out the way I would like" (.54), having the highest loadings on the fourth factor. Although effort rather than outcome seems to characterize several items of Factor 4, this is not a consistent theme. In addition, all of the items in Factor 4 are phrased negatively, which suggests a possible overriding response bias. The final factor, then, is not easily interpretable.

Although the factor analysis did yield a moderately interpretable factor structure, other results militate against assuming the simple factor structure noted above. First, Factor 1 accounted for a disproportionately high percentage of the variance. In addition, 15 items, half of the scale items, loaded greater than $\pm .30$ on at least two and in some cases three factors. The considerable overlap in loadings on the four factors suggests that the factors are not independent. The small sample size and lack of uniformly interpretable factors further limit the validity of a

simple structure. Based on current data, the presence of one general factor is tentatively reasonable. Subsequent factor analyses using larger samples and separate analyses for males and females may prove more conclusive.

Discussion

Results indicate that the GESS has an acceptable test-retest reliability, high internal consistency, and a minimal relationship with social desirability. Preliminary factor analysis did not yield strong evidence of a simple subscale structure. It appears that GESS scores are largely a function of one factor reflecting a sense of general efficacy.

A number of theoretical approaches to depression—learned helplessness (Seligman, 1975), social learning theory (Phares, 1972), and Beck's (Beck, 1967, 1976) model among them—focus on the importance of the depressive's negative cognitive set. To estab-

Table 3

Summary of Correlations Between the GESS and Measures of Internal-External Control of Reinforcement from Other Studies

Scale	Males		Females		Total	
	<i>N</i>	<i>r</i>	<i>N</i>	<i>r</i>	<i>N</i>	<i>r</i>
Strickland (Note 2) (Crandall I-E) ^a						
I-E for positive events	50	.32**	50	.43***		
I-E for negative events	50	-.17	50	.00		
I-E Total	50	.07	50	.26*		
Crepeau (Note 1) (Collins' I-E)						
Difficult/Easy World	67	-.35**	107	-.45***		
Just/Unjust World	67	-.20*	107	-.20*		
Predictable/Unpredictable World	67	-.30**	107	-.36***		
Politically Responsive World	67	-.16	107	-.18		
Personal Control	67	-.42***	107	-.54***		
I-E Total	67	-.48***	107	-.48***		
Koerner (Note 3) (Collins' I-E)						
I-E Total					120	-.41***

Note. GESS = Generalized Expectancy for Success Scale. I-E = internal-external.

^a Crandall's scale is scored in internal direction.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

lish the construct validity of the GESS, an assessment of its scores in relation to measures of depressive cognition is crucial. As expected, the GESS was significantly related to measures of depression, with persons who express high expectancies for success being less likely to report themselves as depressed. This relationship has been further corroborated with samples of college students (see Table 2; Koerner, 1977; Crepeau, Note 1; Strickland, Note 2). The significant negative correlations with the Hopelessness Scale obtained in the present investigation provide further support for the construct validity of the GESS, since Beck et al. (1974) have demonstrated that their scale is a quantitative measure of pessimism. Extreme pessimism or helplessness is also correlated with suicide (Minkoff, Bergman, Beck, & Beck, 1973). In Crepeau's (Note 1) study, retrospective self-report data on the frequency of suicidal ideation among college students correlated negatively and significantly with GESS scores. Thus, low scores on the GESS are related consistently to measures of negative cognition, depressive affect and symptomatology, and suicidal ideation.

Anxiety, frequently a concomitant of depression (Beck, 1967, 1976) and an anticipated correlate of low generalized expectancy for success, related negatively and significantly to GESS scores in both Strickland's (Note 2) and Koerner's (1977) studies. An individual with a low generalized expectancy for success tends to report greater anxiety as measured by the Multiple Affect Adjective Checklist (MAACL; Zuckerman, Lubin, & Robins, 1965). Additionally, Koerner reported a significant negative correlation between scores on the Hostility subscale of the MAACL and GESS scores. As in the female sample of the present study, social desirability was positively correlated with GESS scores for the combined male and female samples of Koerner's study ($r = .29$, $p < .01$). However, after partialing out the effects of social desirability, the relationship between GESS and depression scores remained significant ($r = -.44$, $p < .001$).

Numerous studies have demonstrated a positive relationship between an individual's belief in internal control of reinforcement and successful coping behaviors (Lefcourt, 1976; Phares, 1976, Strickland, 1977; Gilmore,

Note 3). In the present study, high GESS scores were related to internality for female subjects but not for males. Other studies by Strickland (Note 2), Crepeau (Note 1), and Koerner (1977) demonstrated the relationship between GESS and internality for both sexes (see Table 3). Crepeau used the Collins (1974) and Levenson and Miller (1976) subscales derived from the factor analyzed I-E scale to measure locus of control. Crepeau found GESS scores negatively and significantly related to each of the five subscales (difficult/easy world, just/unjust world, predictable/unpredictable world, politically responsive world, and personal control) for female subjects and all but the fourth subscale (politically responsive world) for male subjects (trend, $p < .10$). Using an academic achievement measure that distinguishes between positive and negative events of locus of perceived responsibility (Crandall, Note 4), Strickland reported significant negative correlations between GESS scores and total I-E scores and internality for positive but not negative events for female subjects. For males, GESS scores were negatively and significantly correlated with I-E scores on the positive, but not the negative events, subscale. In summary and as expected, GESS scores and a belief in internal control of reinforcement appear to be related both at a general level and across specific dimensions. These relationships seem somewhat attenuated among male subjects. Discriminant validity is demonstrated by the low and generally insignificant correlations between GESS scores and scores on the Social Desirability Scale and the MAACL Hostility subscale.

Support for the construct validity is provided by Fibel (1976). She investigated the relationship between an individual's generalized expectancy for success, task-specific expectancies for success, and differential responses to a learned helplessness paradigm with college females. Data analysis showed a significant positive correlation between GESS scores and specific expectancies for success in novel and ambiguous situations and relatively lower correlations as specific situational information was acquired. Thus, as postulated, one's specific or immediate expectancy increasingly becomes a function of one's

generalized expectancy as the degree of novelty or ambiguity is amplified.

The choice of a measure of generalized versus specific expectancy for success must be determined by the level of analysis desired. Measures of task-specific expectancies will be of greater predictive utility when the level of analysis is task focused. For example, then, the intent of predicting successful performances on mechanical tasks is better determined by a measure of expectancy tailored to mechanics than by a generalized measure. Similarly, measures of specific expectancies within a single need area, such as academic achievement, are preferable when one's predictive purposes are tied to that need area. A generalized measure will be most useful when the level of analysis is broadly defined as an assessment across need areas and situations or in novel or ambiguous circumstances.

Several cogent issues remain unresolved. The degree of relationship between GESS and measures of other personality variables such as self-esteem has not been investigated. The validity of this instrument for populations other than college students must be demonstrated. Additional factor analyses are needed to support the unidimensionality of the measure. The influence of social desirability on GESS scores while moderate for a measure of a culturally highly valued construct must be taken into account in future work with the scale. Nonetheless, at this point, the GESS appears to be theoretically well-founded, empirically sound, and shows promise of predictive utility. Additionally, its brevity and ease of administration further enhance its value. Anticipated are relationships between the GESS and achievement, assertiveness, risk taking, persuasibility, interpersonal skills, and social competence. Further, a measure of this possibly potent cognitive-mediating variable may have implications for predicting psychological well-being, particularly vis-à-vis depressive symptomatology that is typically characterized by negative expectations and low motivation.

Reference Notes

1. Crepeau, J. *Effects of stressful life events and locus of control on suicidal ideation*. Unpublished

- research, University of Massachusetts, Amherst, 1977.
2. Strickland, B. R. *Aspiration behavior and depression*. Unpublished research, University of Massachusetts, Amherst, 1976.
 3. Gilmore, T. M. *Locus of control as a mediator of adaptive behavior in children and adolescents*. Unpublished article, York University, Toronto, Canada, 1976.
 4. Crandall, V. *An adult measure of intellectual achievement responsibility*. Unpublished scale, 1977.
- ### References
- Beck, A. T. *Depression: Causes and treatment*. Philadelphia: University of Pennsylvania Press, 1967.
- Beck, A. T. *Cognitive therapy and the emotional disorders*. New York: International Universities Press, 1976.
- Beck, A. T., Weissman, A., Lester, D., & Trexler, L. The measurement of pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology*, 1974, 42, 861-865.
- Collins, B. E. Four separate components of the Rotter I-E scale: Belief in a difficult world, a just world, a predictable world, and a politically responsive world. *Journal of Personality and Social Psychology*, 1974, 29, 381-391.
- Crowne, D. P., & Marlowe, D. A. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 1960, 24, 349-354.
- Dickstein, L. S., & Kephart, J. L. Effect of explicit examiner expectancy upon WAIS performance. *Psychological Reports*, 1972, 30, 207-212.
- Feather, N. J. Effects of prior success and failure on expectations of success and subsequent performance. *Journal of Personality and Social Psychology*, 1966, 3, 287-298.
- Feather, N. J., & Saville, M. R. Effects of amount of prior success and failure on expectations of success and subsequent task performance. *Journal of Personality and Social Psychology*, 1967, 5, 226-232.
- Fibel, B. L. Contingency of reinforcement and level of success in a learned helplessness paradigm among college females. Unpublished master's thesis, University of Massachusetts, Amherst, 1976.
- Kaiser, H. F. The Varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, 23, 187-200.
- Koerner, F. E. The effects of depression and sex on aggressive affect and behavior toward the self and toward others. (Doctoral dissertation, University of Massachusetts, 1977). *Dissertation Abstracts International*, 1977, 38, 1887B-1888B. (University Microfilms No. 77-22, 025)
- Lefcourt, H. M. *Locus of control: Current trends in theory and research*. Hillsdale, N.J.: Erlbaum, 1976.
- Levenson, H., & Miller, J. Multidimensional locus of control in sociopolitical activists of conservative and liberal ideologies. *Journal of Personality and Social Psychology*, 1976, 33, 199-208.
- Minkoff, K., Bergman, E., Beck, A. T., & Beck, R. Hopelessness, depression, and attempted suicide. *American Journal of Psychiatry*, 1973, 130, 455-459.
- Phares, E. J. A social learning theory approach to psychopathology. In J. B. Rotter, J. E. Chance, & E. J. Phares (Eds.), *Applications of a social learning theory of personality*. New York: Holt, Rinehart & Winston, 1972.
- Phares, E. J. *Locus of control in personality*. Morristown, N.J.: General Learning Press, 1976.
- Resenthal, R., & Jacobson, L. Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports*, 1966, 19, 115-118.
- Rotter, J. B. *Social learning and clinical psychology*. Englewood Cliffs, N.J.: Prentice-Hall, 1954.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Rotter, J. B. A new scale for the measurement of interpersonal trust. *Journal of Personality*, 1967, 35, 651-665.
- Rotter, J. B. Some problems and misconceptions related to the construct of internal versus external control of reinforcement. *Journal of Consulting and Clinical Psychology*, 1975, 43, 56-67.
- Rotter, J. B., Chance, J. E., & Phares, E. J. *Applications of a social learning theory of personality*. New York: Holt, 1972.
- Seligman, M. E. P. Depression and learned helplessness. In D. Rosenhan & P. London (Eds.), *Theory and research in abnormal psychology*. New York: Holt, Rinehart & Winston, 1975.
- Strickland, B. R. Internal versus external control of reinforcement. In T. Blass (Ed.), *Personality and social behaviors*. Hillsdale, N.J.: Erlbaum, 1977.
- Tyler, B. B. Expectancy for eventual success as a factor in problem solving behavior. *Journal of Educational Psychology*, 1958, 49, 166-172.
- Zuckerman, M., Lubin, B., & Robins, S. Validation of the Multiple Affect Adjective Check List in clinical situations. *Journal of Consulting Psychology*, 1965, 29, 594.
- Zung, W. K. A self-rating depression scale. *Archives of General Psychiatry*, 1965, 12, 63-70.

Appendix

The Hale-Fibel Generalized Expectancy for Success Scale

This is a questionnaire to find out how people believe they will do in certain situations. Each item consists of a 5-point scale and a belief statement regarding one's expectations about events. Please indicate the degree to which you believe the statement would apply to you personally by circling the appropriate number. [1 = highly improbable, 5 = highly probable.] Give the answer that you truly believe *best applies to you* and not what you would like to be true or think others would like to hear. Answer the items *carefully*, but do not spend too much time on any one item. Be sure to find an answer for *every item*, even if the statement describes a situation you presently do not expect to encounter. Answer as if you were going to be in each situation. Also try to respond to each item independently when making a choice; do not be influenced by your previous choices.

In the future I expect that I will

1. find that people don't seem to understand what I am trying to say.
2. be discouraged about my ability to gain the respect of others.
3. be a good parent.
4. be unable to accomplish my goals.
5. have a successful marital relationship.
6. deal poorly with emergency situations.
7. find my efforts to change situations I don't like are ineffective.
8. not be very good at learning new skills.
9. carry through my responsibilities successfully.

10. discover that the good in life outweighs the bad.
11. handle unexpected problems successfully.
12. get the promotions I deserve.
13. succeed in the projects I undertake.
14. not make any significant contributions to society.
15. discover that my life is not getting much better.
16. be listened to when I speak.
17. discover that my plans don't work out too well.
18. find that no matter how hard I try, things just don't turn out the way I would like.
19. handle myself well in whatever situation I'm in.
20. be able to solve my own problems.
21. succeed at most things I try.
22. be successful in my endeavors in the long run.
23. be very successful working out my personal life.
24. experience many failures in my life.
25. make a good impression on people I meet for the first time.
26. attain the career goals I have set for myself.
27. have difficulty dealing with my superiors.
28. have problems working with others.
29. be a good judge of what it takes to get ahead.
30. achieve recognition in my profession.

Received June 20, 1977 ■

Assessing the Impact of Life Changes: Development of the Life Experiences Survey

Irwin G. Sarason, James H. Johnson, and Judith M. Siegel
University of Washington

This article describes the development of a new instrument, the Life Experiences Survey, for the measurement of life changes. It was designed to eliminate certain shortcomings of previous life stress measures and allows for separate assessment of positive and negative life experiences as well as individualized ratings of the impact of events. Several studies bearing on the usefulness of the Life Experiences Survey are presented, and the implications of the findings are discussed.

During recent years, numerous studies have investigated the relationship between life stress and susceptibility to physical and psychological problems. Most of these studies have been based on the assumptions that (a) life changes require adaptation on the part of the individual and are stressful, and (b) persons experiencing marked degrees of life change during the recent past are susceptible to physical and psychiatric problems.

There is considerable evidence that a relationship exists between life stress, operationally defined in terms of self-reported life changes, and physical illness (Dohrenwend & Dohrenwend, 1974b). Rahe and Lind (1971) have reported a relationship between life stress and sudden cardiac death. Theorell and Rahe (1971) and Edwards (1971) have provided data suggestive of a link between life stress and myocardial infarction. Holmes (1970) and Rahe (1968) both found a relationship between life stress and major and minor health

changes, and Wyler, Masuda, and Holmes (1971) have shown that life change is related to seriousness of chronic illness.

There also have been studies of non-health-related correlates of life change that have yielded positive results. For example, significant negative relationships between life stress and academic (Harris, 1973) and teacher (Carranza, 1973) performance have been found. Several researchers have demonstrated a relationship between extent of life changes and psychiatric symptomatology (Dekker & Webb, 1974; Paykel et al., 1969). Vinokur and Selzer (1975) and others (e.g., Constantini, Braun, Davis, & Iervolino, 1973) have also found life stress to be related to the occurrence of depression, anxiety, and tension. A comprehensive review of the life stress literature and a consideration of methodological issues in this area of research has been presented by Rabkin and Struening (1976).

Questions of both a methodological and theoretical nature can be raised concerning present methods of assessing life changes. By far the most widely used instrument in life stress research is the Schedule of Recent Experiences (SRE; Holmes & Rahe, 1967). This is a self-administered questionnaire containing a list of 43 events to which subjects respond by checking those events that they have experienced during the recent past (previous 6 months or 1 year). To determine the scoring weights for specific events, Holmes and Rahe (1967) had a large group of subjects rate

This research was funded by the Organizational Effectiveness Research Program, U.S. Office of Naval Research (Code 452), under Contract N00014-75-C-0905, NR 170-804.

The authors wish to thank Kenneth Hoppe, James Larson, and Karen Lindner who have assisted with data collection for these studies. A note of appreciation is also due Ronald E. Smith for his suggestions during various phases of this project.

Requests for reprints should be sent to Irwin G. Sarason, Department of Psychology NI-25, University of Washington, Seattle, Washington 98195.

each of the 43 items with regard to the amount of social readjustment that the various events required. The item *marriage* (assigned a value of 500) was used as an arbitrary standard or anchor point for making ratings. Mean values were obtained for each of the items. These mean values (divided by the constant of 10) were taken to represent the average amount of social readjustment required by the events. The values, termed *life change units*, when summed yield a total life stress score.

Although the development of the SRE represents a valuable initial attempt at the quantification of the impact of life change, its adequacy has been questioned on several counts (Rabkin & Struening, 1976). The SRE was based on the assumption that life changes per se are stressful regardless of the desirability of the events experienced. Therefore, both desirable and undesirable events are combined in determining the life stress score. On the other hand, several writers have questioned the logic of combining positive and negative events (Brown, 1974; Mechanic, 1975; Sarason, De Monchaux & Hunt, 1975). It has been argued that undesirable events (e.g., death of a close family member) may have a very different, and possibly a more detrimental, effect on individuals than positive events (e.g., outstanding personal achievement). It seems reasonable, therefore, to consider conceptualizing life stress primarily in terms of events that exert negative impacts.

Vinokur and Selzer (1975) have provided information that bears on this issue. These investigators used a specially modified version of the SRE, which yielded separate values for positive and negative life change. Several stress-related measures such as self-ratings of depression, anxiety, and tension were used, as well as measures of aggression, paranoia, and suicidal proclivity. The study provided support for a relationship between life changes and several of these measures but only when using a measure of undesirable events. Positive change was not found to be systematically related to the personality measures. Vinokur and Selzer (1975) concluded that

it seems reasonable to reject the notion that adjustment to change per se is the crucial determinant of life stress and its sequelae. Instead, it appears that

the contribution of life events to psychological impairment is mediated by stress that is evoked by some undesirable aspect of the events rather than by change per se (pp. 333-344).

Similar evidence that psychological difficulties are related to undesirable, but not desirable events has been provided by Mueller, Edwards, and Yarvis (1977). It would seem necessary to take this desirability-undesirability dimension into account in the assessment of life change.

Even though it might be advisable to categorize events as being desirable or undesirable for purposes of assessment, there are some difficulties with this approach. For example, events may vary in terms of their desirability depending on the circumstances and perceptions of the respondent. To illustrate, "pregnancy" may be a highly desirable event for a woman who wants a child, but it may be viewed as quite undesirable by an unwed teenager. Given the fact that individuals perceive events differently, it is somehow important to individualize ratings of the desirability of the events that they experience.

A related issue concerns the quantification of life changes. Because individuals vary in how they are affected by events, the values derived from group ratings (such as those used with the SRE) may not accurately reflect the impact that events have on particular individuals. Problems inherent in applying group-derived values to individual cases become obvious when it is noted that certain classes of events listed in the SRE can be quite ambiguous. For instance, if a subject responds to the item *major change in financial status*, it is uncertain if the response refers to a major change in a positive or negative direction. It is not clear that the life change unit associated with major change in financial status is as appropriate to the person who has recently become bankrupt as to the person who has recently inherited a large sum of money. Thus, even though life change units do seem to provide a quantitative measure of overall life change, in some cases, they may not reflect the actual amount of stress resulting from the experiencing of specific events. Findings bearing on this issue have recently been reported by Yamamoto and Kinney (1976). These investigators found life

stress scores, based on self-ratings of the stressfulness of events, to be better predictors than scores derived by using mean adjustment ratings similar to those used with the SRE.

Bearing in mind the methodological issues mentioned above, it would appear that a measure of life stress should possess three characteristics. First, it should include a list of events experienced with at least some degree of frequency in the population being investigated. Second, it should allow for ratings, by respondents themselves, of the desirability or undesirability of the events. Third, it should allow for individualized ratings of the personal impact of the events experienced. The present article describes a new measure of life stress, the Life Experiences Survey (LES), constructed according to these guidelines and describes the results of several studies bearing on its usefulness.

Development of the LES

The LES is a 57-item self-report measure that allows respondents to indicate events that they have experienced during the past year. The scale has two portions: Section 1, designed for all respondents, contains a list of 47 specific events plus three blank spaces in which subjects can indicate other events that they may have experienced. The events listed in Section 1 refer to life changes that are common to individuals in a wide variety of situations. The 10 events listed in Section 2 are designed primarily for use with students, but they can be adapted for other populations. Section 2 deals specifically with changes experienced in the academic environment. Section 1 is appropriate for use with subjects drawn from the general population, whereas both sections are relevant to a student population. (In the present research, responses to items of Sections 1 and 2 were combined in deriving life change scores as this research was conducted with college students.)

The LES items were chosen to represent life changes frequently experienced by individuals in the general population. Many of the items are based on existing life stress measures, particularly the SRE. Others were included because they were judged to be events

that occur frequently and that potentially might exert a significant impact on the lives of persons experiencing them. Thirty-four of the events listed in the LES are similar in content to those found in the SRE (Holmes & Rahe, 1967). In the construction of the present scale, however, certain items were made more specific. For example, the SRE contains the item *pregnancy*, which may be endorsed by women but perhaps not by a man whose wife or girlfriend has become pregnant. The present scale allows both men and women to endorse the item of pregnancy in the following manner: *Female: Pregnancy; Male: Wife's/girlfriend's pregnancy*. The SRE includes the item *Wife begins or stops work*, an item that fails to assess the impact on women whose husbands begin or cease working. The present scale lists two items: *Married male: Change in wife's work outside the home (beginning work, ceasing work, changing to a new job, etc.)*, and *Married female: Change in husband's work (loss of job, beginning a new job, etc.)* Examples of events not listed in the SRE but included here are male and female items dealing with abortion and more general items such as serious injury or illness of close friend, engagement, breaking up with boyfriend/girlfriend, and so forth. Nine of the 10 school-related items are unique to the LES. Finally, some of the events from the SRE thought to be of relatively little consequence (e.g., Christmas, vacation, etc.) were not included, and certain other events were reworded to simplify responding.

The format of the LES calls for subjects to rate separately the desirability and impact of events that they have experienced. Thus, they are asked to indicate those events experienced during the past year (0-6 months or 7 months-1 year)¹ as well as (a) whether they viewed the event as being positive or negative and (b) the perceived impact of the particular event on their life *at the time of occurrence*. Ratings are on a 7-point scale ranging from extremely negative (-3) to ex-

¹ Although the LES provides for the assessment of life change occurring during two 6-month intervals, all analyses to date have involved change scores based on the entire preceding 12-month time period.

tremely positive (+3). Summing the impact ratings of those events designated as positive by the subject provides a *positive change score*. A *negative change score* is derived by summing the impact ratings of those events experienced as negative by the subject. By adding these two values, a *total change score* can be obtained, representing the total amount of rated change (desirable and undesirable) experienced by the subject during the past year. Although the findings cited earlier (Mueller et al., 1977; Vinokur & Selzer, 1975) suggest that this total change score might be less predictive of health-related variables than an index of negative change, this measure was used in the present research to provide further information concerning the relationships between negative change, change per se, and stress-related dependent variables. (The LES is presented in the Appendix.)

For any new instrument it is necessary to obtain certain kinds of information. Normative data should be provided that include information about the effects of demographic variables (e.g., sex). Evidence should also be presented concerning the instrument's stability over time and correlations with relevant dependent measures. Finally, in the case of self-report scales, it should be demonstrated that measures derived from the instrument do not simply reflect the effects of response sets such as the tendency to "fake good." The instrument's scores should not be highly correlated with factors such as social desirability.

Normative Data and an Examination of Sex Differences

The first study undertaken with the LES obtained general information concerning the responses of college students to the instrument and investigated the possibility of differences in response due to sex.

The LES was administered in class to 345 students enrolled in introductory psychology courses at the University of Washington during the fall quarter of 1975. Values were obtained for positive, negative, and total life change scores. Means and standard deviations were derived separately for males ($n = 174$) and females ($n = 171$) on each of these mea-

Table 1
Means and Standard Deviations of Male and Female Respondents on the Life Experiences Survey (LES)

LES score	Males ^a		Females ^b	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Positive	9.74 6.87	8.07 5.97	9.57 6.71	6.66 5.51
Negative	6.22 4.66	6.28 4.36	7.04 5.64	7.90 6.43
Total	15.97 11.53	11.08 8.01	16.61 12.35	10.23 8.82

Note. In each case figures in top rows are derived from responses to Parts 1 and 2 combined. Figures in the bottom rows are derived from Part 1 only.

^a $n = 174$.

^b $n = 171$.

asures, and tests for sex differences were conducted.² Data from Section 1 and Sections 1 and 2 combined are presented separately in Table 1. The table shows that there were no significant differences between males and females on any of the three life change measures. It can also be seen that the life change scores of this sample of college students are generally low. Higher values might well have been obtained if subjects from the general population had been surveyed. Finally, it can be noted that the results of this and a number of other studies with the LES have shown that the positive and negative life change scores are essentially uncorrelated.

Reliability of the LES

Two test-retest reliability studies of the LES have been conducted. Both involved subjects drawn from undergraduate psychology courses with a 5- to 6-week time interval between test and retest. There were 34 subjects in the first study and 58 in the second. Responses were scored for positive, negative,

² Data concerning the mean ratings of these events; the frequency of endorsement of various events; and percentile ranks for positive, negative, and total change scores can be obtained without charge from the authors.

and total life changes in each case. Pearson product-moment correlations were computed to determine the relationships between scores obtained at the two testings. Test-retest correlations for the positive change score were .19 and .53 ($p < .001$). The reliability coefficients for the negative change score were .56 ($p < .001$) and .88 ($p < .001$). The coefficients for the total change score were .63 ($p < .001$) and .64 ($p < .001$).

Although the findings of the two studies reported here vary to some extent, perhaps due to the relatively small sample sizes, they suggest that LES is a moderately reliable instrument especially when the negative and total change scores are considered.³ It should be noted that test-retest reliability coefficients found with instruments of this type are likely to underestimate the reliability of the measure. That is, with a time interval of 5-6 weeks, subjects may actually experience a variety of events, both positive and negative, that may be reflected in responses given at the time of retesting. As these changes reflect the actual occurrence of life changes, rather than simply inconsistencies in reporting, it would be inappropriate to consider the total variability in responding as error. As subjects generally seem to report somewhat higher levels of positive than negative change on the LES, it seems possible that the lower reliability estimates found with the positive change measure may be due, in part, to the greater likelihood of positive changes occurring within the time interval between test and retest.

Correlates of the LES

To the extent that the LES measures life stress, its scores should correlate with relevant personality indices. Further, an analysis of the correlational patterns should provide information concerning whether life stress is more usefully conceptualized in terms of negative life change or total life change.

Anxiety, Academic Achievement, Social Desirability, and the LES

A group of 100 male and female students drawn from introduction to personality

courses were administered the LES, the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970), and a short form of the Marlowe-Crowne Social Desirability Scale (Strahan & Gerbasi, 1972). Academic transcripts were available for 75 of these students. The correlations among life change scores, anxiety, and grade point average (GPA) are presented in Table 2.

Inspection of these correlations shows that the total and negative change scores correlate significantly and in a positive direction with state and trait anxiety, whereas the positive change score is not significantly related to either measure. Tests for significance of the difference between correlations suggested that positive and negative change scores differ significantly in their correlations with state anxiety ($p < .01$). Although negative change scores were significantly correlated with trait anxiety and positive change scores were not, the difference between these correlations was not significant. Significant correlations between negative change and anxiety have also been found in data collected as part of two other investigations. For a sample of naval personnel ($N = 76$), correlations of .46 ($p < .001$) and .40 ($p < .001$) were found with state and trait anxiety, respectively. With college students ($N = 82$), a correlation of .24 ($p < .05$) has also been found between negative change and anxiety as measured by the Multiple Affect Adjective Checklist (Zuckerman & Lubin, 1965).

With regard to GPA, positive, negative, and total change scores were all found to be negatively correlated with GPA. Even though the correlation between positive change and GPA was smaller than the correlations between the negative and total change scores and this measure, the differences between these correlations were not significant. These results are

³ It should be noted that in addition to the two reliability studies reported here, reliability data are available on one additional small sample of 12 subjects who took the LES on two different occasions as part of another investigation. The time interval between test and retest was 8 weeks. Here, reliability coefficients of .61 ($p < .05$), .72 ($p < .01$), and .82 ($p < .001$) were obtained for positive, negative, and total change scores, respectively.

Table 2
Correlations Between Life Change Scores, Anxiety, and Academic Achievement

LES Life change scores	Anxiety		Grade point average ^b
	Trait ^a	State ^a	
Positive	.04	.03	-.21
Negative	.29**	.46***	-.38***
Total	.24*	.37***	-.40***
Balance (negative - positive events)	-.21*	-.36***	.18

^ae. LES = Life Experiences Survey.

^b = 97.

^c = 73.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

consistent with other studies that have found significant relationships between life stress (assessed by other measures) and measures of anxiety (Constantini et al., 1973) and academic achievement (Carranza, 1973).

As it seemed reasonable that the effects of positive change might, in part, ameliorate the stress produced by negative experiences, a balance or subtractive score (negative - positive) was also computed for each subject and was correlated with the dependent measures. As can be seen in Table 2, in no case was the balance score more predictive than the positive change score alone, although differences between correlations were not significant. These results are similar to those reported by Mueller et al. (1977) and Vinokur and Selzer (1975), who have found such a balance score to be less predictive of stress-related variables than measures of negative life change.

The relationships between life change scores and the social desirability measure were non-significant. Correlations between positive, negative, and total change scores and social desirability were $-.05$, $.05$, and $.01$, respectively. This suggests that responses to the LES are relatively free from the influence of social desirability response bias.

Personal Maladjustment and the LES

To determine the relationship between life stress and personal maladjustment, the LES and the Psychological Screening Inventory

(PSI) were administered to 75 male and female volunteers drawn from introduction to personality courses at the University of Washington.

The PSI (Lanyon, 1970, 1973) is a 130-item true-false inventory that yields scores on five subscales: Alienation (Al), Social Nonconformity (Sn), Discomfort (Di), Expression (Ex), and Defensiveness (De). The Al scale was designed for "assessing similarity to psychiatric patients," and the Sn scale, for "assessing similarity to incarcerated prisoners." The Di scale appears to be a measure of neuroticism, the Ex scale is a measure of the introversion-extraversion dimension, and the De scale is a measure of test-taking attitude.

Correlations between positive, negative, and total life change scores and the five PSI scales are presented in Table 3. The table shows that negative life change is significantly related to scores on the Sn and Di scales. These findings suggest a relationship between negative change, as assessed by the LES, and certain types of personal maladjustment. Although two PSI scales were correlated with negative change only, the PSI Ex scale was found to correlate significantly with the positive change score. Thus, it would appear that extraverted individuals experience greater degrees of positive change than do introverted persons. The results obtained here are similar to those obtained by Constantini et al. (1973) in their investigation correlating life stress scores, derived from the Holmes and Rahe (1967) scale, with PSI scores. The fact that

Table 3
Correlations Between Life Change and PSI Scores

Life change	PSI				
	Al	Sn	Di	Ex	De
Positive	.14	.03	-.07	.28**	.06
Negative	-.10	.20*	.23*	-.02	-.16
Total	.03	.15	-.10	.18	-.06

Note. PSI = Psychological Screening Inventory; Al = Alienation; Sn = Social Nonconformity; Di = Discomfort; Ex = Expression; De = Defensive-ness.

* $p < .05$.

** $p < .02$.

in the present study PSI measures of personal maladjustment as well as certain of the measures considered earlier (e.g., anxiety) correlate with negative but not with positive change provide further support for the notion that life stress may best be conceptualized in terms of negative change.

Depression, Locus of Control, and the LES

Scores on the LES, the Beck Depression Inventory (Beck, 1967), and the Internal-External (I-E) Locus of Control Scale (Rotter, 1966) were obtained for a sample of 64 (34 males 30 females) college students drawn from undergraduate psychology courses. Correlations between life change scores and these two measures are presented in Table 4. The table reveals a significant relationship between negative change and scores on the Beck Depression Inventory, which is consistent with evidence presented by Vinokur and Selzer (1975), who found negative change to be related to self-ratings of depression. An additional finding of interest is that individuals who report having experienced high levels of negative change appear to be more externally oriented, perceiving themselves as being less capable of exerting control over reinforcement contingencies in their environment.

A Study of Counseling Center Clients

In addition to the findings presented above, life change scores have also been obtained

from a group of students receiving treatment at a university counseling center for psychological problems. Based on earlier findings of a relationship between negative life change and measures of personal maladjustment, it was predicted that this group would differ from a randomly selected group of college students in their negative change scores but not in terms of positive change. The counseling center sample consisted of 18 students (16 females and 2 males). For purpose of comparison, LES records of 18 (16 females and 2 males) students were selected at random from protocols obtained from students enrolled in introduction to personality courses at the University of Washington. (Undergraduates at all academic levels are enrolled in these courses.) Mean positive, negative, and total change scores for these two groups are presented in Table 5.

No significant differences were obtained for the positive and total change scores. The counseling center clients did, however, display significantly higher negative change scores than did the comparison group, $t(34) = 2.21$, $p < .05$. In order to rule out the possibility that these findings are unique to the random sample selected for comparison, a second comparison group ($n = 18$) was randomly drawn from the completed LES protocols of introductory psychology students. Again, significance between group means was found for negative, $t(34) = 2.89$, $p < .01$, but not for positive or total life change. These findings provide additional support for a relationship between negative life change as assessed by the LES and problems of a psychological nature.

Table 4
Correlations Between Life Change, Depression, and Locus of Control

Life change score	Beck depression	Locus of control
Positive	-.15	-.05
Negative	.24*	.32**
Total	.06	.17

* $p < .05$.

** $p < .02$.

Table 5
Life Change Scores for Normals and Counseling Center Clients

Group	Change					
	Positive		Negative		Total	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Normals	10.55	8.26	9.61	9.59	20.16	11.48
Counseling center	8.33	5.83	16.61	9.37	24.94	10.91

Note. $n = 18$ for both groups.

A Comparison of the LES and SRE Approaches as Measures of Life Change

If the LES represents an improvement over the SRE, it should be possible to demonstrate that measures derived from the LES are more highly related to relevant dependent measures than are SRE scores. Further analyses of some of the data already reported, along with analyses of additional data, were undertaken to provide some basis for comparing these two indices of life stress. The comparisons were accomplished by scoring only the 34 items of the LES that are common to the Holmes and Rahe (1967) measure. These items were scored to yield four measures. Three of these measures were LES positive, negative, and total life change scores derived in the manner described earlier. A fourth measure was derived by applying the life change units used with the SRE to each of the 34 items. It was thus possible to derive a measure comparable to the SRE based on responses to these events. Although these measures were based on 34 rather than the entire 43 items of the Holmes and Rahe scale, it was felt that they would provide an adequate basis for comparing the LES and SRE scoring procedures. Based on the findings reported earlier, it was predicted that the LES negative change score would be more predictive of dependent measures than would the Holmes and Rahe measure. No predictions were made regarding the LES positive and total change scores.

In one comparative study, 69 female subjects from undergraduate human sexuality courses were given the LES, the Beck Depression Inventory, and the State-Trait Anx-

iety Inventory. The four life change measures were derived as outlined above. One somewhat surprising finding was that no significant correlations were found between any of the four life change measures and anxiety. Given the rather consistent finding of a relationship between negative change and anxiety reported earlier, these results might best be attributed to the rather select nature of the sample studied. Significant findings were, however, obtained for correlations with the Beck Depression Inventory. Correlations between positive, negative, and total LES scores and depression were .02, .37 ($p < .01$), and .24 ($p < .05$), respectively. The correlation between the life change unit score, similar to that used with the SRE, and depression was .17 (ns). The difference between the correlations obtained with the LES negative change score and the Holmes and Rahe score was significant, $t(66) = 2.31$, $p < .05$.

A second comparative study of the LES and SRE measures concerned the relationship between these measures and the scores on the PSI. As in the original analysis (which included the entire LES), two PSI adjustment measures were found to be significantly correlated with life change when only 34 items were scored, Sn and Di (neuroticism). Correlations between change scores and these measures are presented in Table 6. As can be seen, the LES negative change scores correlated significantly with both measures of adjustment (Sn and Di), whereas no significant relationships were found between these two measures and the life change unit score. Although the differences between these correlations did not reach statistical significance, the pattern of results does seem to support the

Table 6
Correlations Between LES Change Scores, Life Change Unit Scores (34 items), and Psychological Screening Inventory (PSI) Scale Scores

LES Life Change score	PSI	
	Sn	Di
Positive	.02	-.04
Negative	.26*	.25*
Total	.18	.12
Unit	.14	.15

Note. LES = Life Experiences Survey; Sn = Social Nonconformity; Di = Discomfort.

* $p < .05$.

superiority of the LES measure of negative change.

Discussion

The results of the studies reported here suggest that the LES may be a useful research and, perhaps also, clinical tool. They indicate that negative and total change scores, derived from this scale, are reasonably reliable over a 5- to 6-week time interval, although the positive change score appears to be less stable. Support for the usefulness of the scale is provided by the findings showing that the negative life change score is significantly related to a number of stress-related dependent measures. In addition, scale responses appear to be relatively free from social desirability biases, and the measure is capable of differentiating college students who have sought help for adjustment problems from those who have not.

Other results also suggest that the LES possesses certain advantages over the SRE as an instrument for assessing life stress. These advantages relate particularly to the important distinction between desirable and undesirable change made by the LES. The results show that positive and negative life change scores exhibit different patterns of relationships with relevant dependent measures. It can be noted that there was not a single case in which both positive and negative change scores were significantly correlated with the

same dependent measures. This suggests that the separate assessment of positive and negative change by the LES represents a step forward in assessing relationships between life changes and diverse dependent measures. It seems possible that life stress is most accurately conceptualized in terms of negative life changes rather than in terms of positive or total change. Our findings and those reported by others suggest that it is the negative change measure that should be used if one's purpose is to determine degree of "life stress."

Although the results reported here emphasize the role of negative change, it should be pointed out that the failure to find significant correlations between positive change and the dependent measures may be related to the lower reliability of the positive change score rather than to the unstressful nature of positive life change. The findings of Mueller et al. (1977) and Vinokur and Selzer (1975), which are consistent with the present results, however, support conclusions emphasizing the importance of negative life changes.

A major consideration in the assessment of life stress concerns the nature of the relationships obtained between life change scores and stress-related dependent variables. One might question, for example, whether relationships such as those reported in this article and found elsewhere in the literature reflect the effects of life stress on individuals or simply reflect the effects of specific variables on the reporting of life change. Regarding life stress research in general, one might also question whether persons experiencing high levels of life stress are actually more susceptible to the development of physical and/or psychological problems or whether persons who already manifest such difficulties are more prone to experience life change. Thus, the directionality of the relationships obtained in life stress studies is often unclear. This makes it difficult to draw firm cause-effect conclusions. Although authors such as Brown (1972) have made a strong case for the causal role of life stress, and even though most research in the area seems to be based on the assumption that change plays a causal role, definitive answers regarding cause-effect relationships must ultimately come from longitudinal stud-

ies that are more complex than those typically found in the life stress literature.

Although based on available research findings, it is not possible to resolve this directionality issue even though some data are available regarding the degree to which life stress scores may themselves be influenced by the psychological state of the respondent at the time of testing. In a recent study by Siegel, Johnson, and Sarason (Note 1), the effects of an experimentally induced depressive state on responding to the LES was investigated. The subjects, who had previously completed the LES, were randomly assigned to one of three experimental conditions: neutral, elation, and depression. By using an affect induction procedure developed by Velten (1968), it was possible to induce transient states of elation and depression in these subjects. Subjects were then given the LES a second time. Although a manipulation check indicated that the affect induction procedure did result in elation and depression in the two experimental groups, these mood states had no effect on the number of life changes reported or on any of the LES scores. These results suggest that the significant correlations between the LES and depression do not result from the effects of the depressive mood state on responding to the LES. These results might be interpreted as being consistent with the notion that a causal relationship exists between negative events and depression. However, additional data are needed to draw firm conclusions. (Although mood state does not influence responding per se, depressed individuals as a result of their condition may actually experience more negative changes, thus resulting in a correlation between change and depression.) The results do suggest, however, that responses to the LES are not unduly influenced by the mood state of the respondent.

Finally, in considering the assessment of life change and its effect on individuals, it would seem necessary to take into account the role of variables in addition to life stress. For example, it may be noted that even though significant relationships between change scores and dependent measures were found in this research, the magnitude of the correlations was in most instances low, sug-

gesting that life stress accounts for a relatively small proportion of the variance reflected in the measures. This finding of significant but low correlations is consistent with the results of other life stress studies. It thus seems appropriate to question whether these findings reflect the inadequacy of present life stress measures or if it is, in fact, reasonable to expect such measures to correlate highly with stress-related variables. Dohrenwend and Dohrenwend (1974a) have pointed out that it is likely that the effects of life stress differ from person to person depending on their individual characteristics. Some persons may be greatly affected by even moderate levels of life change, whereas others may be affected very little by relatively high levels. If this is the case, it may not be unreasonable to expect correlations of the low magnitude that have typically been obtained. Perhaps we can expect to find stronger relationships only as variables determining the effects of life change are taken into account.

Unfortunately, relatively little research has been directed toward investigating the role of moderator variables, although the research that has been conducted is provocative. Nuckolls, Cassel, and Kaplan (1972) investigated the relationships between life stress and pregnancy and birth complications. No significant relationships were found among these variables when all subjects were considered. However, when mothers were divided into those who displayed high and low levels of "psychosocial assets," significant results were obtained. Subjects showing high levels of both life change and psychosocial assets (support systems in their environment) did not show evidence of increased complications. Those who displayed high levels of life change and low levels of psychosocial assets did have an increased frequency of such complications.

The importance of moderator variables has also been suggested by the results of a study conducted by Johnson and Sarason (in press) in which the relationships among life change and measures of anxiety and depression were examined as a function of locus of control orientation (Rotter, 1966). It was predicted that a relationship between negative change and depression and anxiety would be found

for externally oriented subjects (who presumably see themselves as having little control over environmental events) but not for internally oriented subjects (who tend to perceive themselves as capable of exerting control over environmental events). The results were in line with this prediction, thus suggesting that life stress may affect individuals differently depending on the degree of their perceived control over events. In one other study, Smith, Johnson, and Sarason (1978) found the relationship between life stress and a measure of psychological adjustment to vary as a function of subjects' scores on a measure of sensation seeking (Zuckerman, Kolin, Price, & Zoob, 1964). Thus, the effects of life stress may also be mediated by self-reported "optimal level of stimulation."

It would appear, then, that one's perception of control over environmental events, sensation-seeking status, and degree of psychosocial assets may all mediate the effects of life stress. It seems likely that there are also other individual difference variables that moderate the effects of life changes, and research designed to identify them is needed. The LES, which possesses sufficient reliability and correlates with a variety of relevant dependent measures, could be used in studies aimed at identifying moderator variables and their effects. The format of the LES allows for the individualized rating of the impact of events plus the availability of separate measures of positive and negative change. This makes it especially appropriate for use in future research concerning how people deal with the stresses and strains of modern life.

Reference Note

1. Siegel, J. M., Johnson, J. H., & Sarason, I. G. *Mood states and the reporting of life change*. Unpublished manuscript, University of Washington, 1978.

References

- Beck, A. T. *Depression: Clinical, experimental, and theoretical aspects*. New York: Harper & Row, 1967.
- Brown, G. W. Life-events and psychiatric illness: Some thoughts on methodology and causality. *Journal of Psychosomatic Research*, 1972, 16, 311-320.
- Brown, G. W. Meaning, measurement, and stress of life events. In B. S. Dohrenwend & B. P. Dohrenwend (Eds.), *Stressful life events: Their nature and effects*. New York: Wiley, 1974.
- Carranza, E. A study of the impact of life changes on high school teacher performance in the Lansing school district as measured by the Holmes and Rahe Schedule of Recent Experiences. (Doctoral dissertation, Michigan State University, 1972). *Dissertation Abstracts International*, 1973, 33, 4996A-4997A. (University Microfilms No. 73-5340)
- Constantini, A. F., Braun, J. R., Davis, J., & Iervolino, A. Personality and mood correlates of Schedule of Recent Experience scores. *Psychological Reports*, 1973, 32, 416-418.
- Dekker, D. J., & Webb, J. T. Relationships of the social readjustment rating scale to psychiatric patient status, anxiety, and social desirability. *Journal of Psychosomatic Research*, 1974, 18, 125-130.
- Dohrenwend, B. S., & Dohrenwend, B. P. Overview and prospects for research on stressful life events. In B. S. Dohrenwend & B. P. Dohrenwend (Eds.), *Stressful life events: Their nature and effects*. New York: Wiley, 1974. (a)
- Dohrenwend, B. S., & Dohrenwend, B. P. *Stressful life events: Their nature and effects*. New York: Wiley, 1974. (b)
- Edwards, M. K. *Life crises and myocardial infarction*. Unpublished master's thesis, University of Washington, 1971.
- Harris, P. W. The relationship of life change to academic performance among selected college freshmen at varying levels of college readiness. (Doctoral dissertation, East Texas State University, 1972). *Dissertation Abstracts International*, 1973, 33, 6665A-6666A. (University Microfilms No. 73-14285)
- Holmes, T. H., & Rahe, R. H. The social readjustment rating scale. *Journal of Psychosomatic Research*, 1967, 11, 213-218.
- Holmes, T. S. *Adaptive behavior and health change*. Unpublished medical thesis, University of Washington, Seattle, 1970.
- Johnson, J. H., & Sarason, I. G. Life stress, depression and anxiety: Internal-external control as a moderator variable. *Journal of Psychosomatic Research*, in press.
- Lanyon, R. I. Development and validation of a psychological screening inventory. *Journal of Consulting and Clinical Psychology*, 1970, 35, 1-24.
- Lanyon, R. I. *Psychological Screening Inventory manual*. Goshen, N.Y.: Research Psychologists Press, 1973.
- Mechanic, D. Some problems in the measurement of stress and social readjustment. *Journal of Human Stress*, 1975, 1, 43-48.
- Mueller, D. P., Edwards, D. W., & Yarvis, R. M. Stressful like events and psychiatric symptomatology: Change or undesirability? *Journal of Health and Social Behavior*, 1977, 18, 307-317.

- Nuckolls, K. B., Cassel, J., & Kaplan, B. H. Psychosocial assets, life crisis and the prognosis of pregnancy. *American Journal of Epidemiology*, 1972, 95, 431-441.
- Paykel, E. S., et al. Life events and depression. *Archives of General Psychiatry*, 1969, 21, 753.
- Rabkin, J. G., & Struening, E. L. Life events, stress, and illness. *Science*, 1976, 194, 1013-1020.
- Rahe, R. H. Life-change measurement as a predictor of illness. *Proceedings of the Royal Society of Medicine*, 1968, 61, 1124-1126.
- Rahe, R. H., & Lind, E. Psychosocial factors and sudden cardiac death: A pilot study. *Journal of Psychosomatic Research*, 1971, 15, 19-24.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Sarason, I. G., De Monchaux, C., & Hunt, T. Methodological issues in the assessment of life stress. In L. Levi (Ed.), *Emotions—Their parameters and measurement*. New York: Raven Press, 1975.
- Smith, R. E., Johnson, J. H., & Sarason, I. G. Life change, the sensation seeking motive, and psychological distress. *Journal of Consulting and Clinical Psychology*, 1978, 46, 348-349.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *Manual for the State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Strahan, R., & Gerbasi, K. C. Short homogeneous versions of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 1972, 28, 191-193.
- Theorell, T., & Rahe, R. H. Psychosocial factors and myocardial infarction. I: An inpatient study in Sweden. *Journal of Psychosomatic Research*, 1971, 15, 25-31.
- Velten, E. A. A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 1968, 6, 473-482.
- Vinokur, A., & Selzer, M. L. Desirable versus undesirable life events: Their relationship to stress and mental distress. *Journal of Personality and Social Psychology*, 1975, 32, 329-337.
- Wyler, A. R., Masuda, M., & Holmes, T. H. Magnitude of life events and seriousness of illness. *Psychosomatic Medicine*, 1971, 33, 115-122.
- Yamamoto, K. J., & Kinney, O. K. Pregnant women's ratings of different factors influencing psychological stress during pregnancy. *Psychological Reports*, 1976, 39, 203-214.
- Zuckerman, M., Kolin, E. A., Price, L., & Zoob, I. Development of a sensation seeking scale. *Journal of Consulting Psychology*, 1964, 26, 250-260.
- Zuckerman, M., & Lubin, B. *Manual for The Multiple Affect Adjective Check List*. San Diego, Calif.: Educational and Industrial Testing Service, 1965.

Appendix

The Life Experiences Survey

Listed below are a number of events which sometimes bring about change in the lives of those who experience them and which necessitate social readjustment. Please check those events which you have experienced in the recent past and indicate the time period during which you have experienced each event. Be sure that all check marks are directly across from the items they correspond to.

Also, for each item checked below, please indicate the extent to which you viewed the event as having either a positive or negative impact on your life at the time the event occurred. That is, indicate the type and extent of impact that the event had. A rating of -3 would indicate an extremely negative impact. A rating of 0 suggests no impact either positive or negative. A rating of +3 would indicate an extremely positive impact.

Section 1

	0 to 6 mo	7 mo to 1 yr	extremely negative	moderately negative	somewhat negative	no impact	slightly positive	moderately positive	extremely positive
1. Marriage			-3	-2	-1	0	+1	+2	+3
2. Detention in jail or comparable institution			-3	-2	-1	0	+1	+2	+3
3. Death of spouse			-3	-2	-1	0	+1	+2	+3
4. Major change in sleeping habits (much more or much less sleep)			-3	-2	-1	0	+1	+2	+3

	0 to 6 mo	7 mo to 1 yr	extremely negative	moderately negative	somewhat negative	no impact	slightly positive	moderately positive	extremely positive
5. Death of close family member:									
a. mother			-3	-2	-1	0	+1	+2	+3
b. father			-3	-2	-1	0	+1	+2	+3
c. brother			-3	-2	-1	0	+1	+2	+3
d. sister			-3	-2	-1	0	+1	+2	+3
e. grandmother			-3	-2	-1	0	+1	+2	+3
f. grandfather			-3	-2	-1	0	+1	+2	+3
g. other (specify)			-3	-2	-1	0	+1	+2	+3
6. Major change in eating habits (much more or much less food intake)			-3	-2	-1	0	+1	+2	+3
7. Foreclosure on mortgage or loan			-3	-2	-1	0	+1	+2	+3
8. Death of close friend			-3	-2	-1	0	+1	+2	+3
9. Outstanding personal achievement			-3	-2	-1	0	+1	+2	+3
10. Minor law violations (traffic tickets, disturbing the peace, etc.)			-3	-2	-1	0	+1	+2	+3
11. <i>Male</i> : Wife/girlfriend's pregnancy			-3	-2	-1	0	+1	+2	+3
12. <i>Female</i> : Pregnancy			-3	-2	-1	0	+1	+2	+3
13. Changed work situation (different work responsibility, major change in working conditions, working hours, etc.)			-3	-2	-1	0	+1	+2	+3
14. New job			-3	-2	-1	0	+1	+2	+3
15. Serious illness or injury of close family member:									
a. father			-3	-2	-1	0	+1	+2	+3
b. mother			-3	-2	-1	0	+1	+2	+3
c. sister			-3	-2	-1	0	+1	+2	+3
d. brother			-3	-2	-1	0	+1	+2	+3
e. grandfather			-3	-2	-1	0	+1	+2	+3
f. grandmother			-3	-2	-1	0	+1	+2	+3
g. spouse			-3	-2	-1	0	+1	+2	+3
h. other (specify)			-3	-2	-1	0	+1	+2	+3
16. Sexual difficulties			-3	-2	-1	0	+1	+2	+3
17. Trouble with employer (in danger of losing job, being suspended, demoted, etc.)			-3	-2	-1	0	+1	+2	+3
18. Trouble with in-laws			-3	-2	-1	0	+1	+2	+3
19. Major change in financial status (a lot better off or a lot worse off)			-3	-2	-1	0	+1	+2	+3
20. Major change in closeness of family members (increased or decreased closeness)			-3	-2	-1	0	+1	+2	+3
21. Gaining a new family member (through birth, adoption, family member moving in, etc.)			-3	-2	-1	0	+1	+2	+3
22. Change of residence			-3	-2	-1	0	+1	+2	+3
23. Marital separation from mate (due to conflict)			-3	-2	-1	0	+1	+2	+3
24. Major change in church activities (increased or decreased attendance)			-3	-2	-1	0	+1	+2	+3

	0 to 6 mo	7 mo to 1 yr	extremely negative	moderately negative	somewhat negative	no impact	slightly positive	moderately positive	extremely positive
25. Marital reconciliation with mate			-3	-2	-1	0	+1	+2	+3
26. Major change in number of arguments with spouse (a lot more or a lot less arguments)			-3	-2	-1	0	+1	+2	+3
27. <i>Married male</i> : Change in wife's work outside the home (beginning work, ceasing work, changing to a new job, etc.)			-3	-2	-1	0	+1	+2	+3
28. <i>Married female</i> : Change in husband's work (loss of job, beginning new job, retirement, etc.)			-3	-2	-1	0	+1	+2	+3
29. Major change in usual type and/or amount of recreation			-3	-2	-1	0	+1	+2	+3
30. Borrowing more than \$10,000 (buying home, business, etc.)			-3	-2	-1	0	+1	+2	+3
31. Borrowing less than \$10,000 (buying car, TV, getting school loan, etc.)			-3	-2	-1	0	+1	+2	+3
32. Being fired from job			-3	-2	-1	0	+1	+2	+3
33. <i>Male</i> : Wife/girlfriend having abortion			-3	-2	-1	0	+1	+2	+3
34. <i>Female</i> : Having abortion			-3	-2	-1	0	+1	+2	+3
35. Major personal illness or injury			-3	-2	-1	0	+1	+2	+3
36. Major change in social activities, e.g., parties, movies, visiting (increased or decreased participation)			-3	-2	-1	0	+1	+2	+3
37. Major change in living conditions of family (building new home, remodeling, deterioration of home, neighborhood, etc.)			-3	-2	-1	0	+1	+2	+3
38. Divorce			-3	-2	-1	0	+1	+2	+3
39. Serious injury or illness of close friend			-3	-2	-1	0	+1	+2	+3
40. Retirement from work			-3	-2	-1	0	+1	+2	+3
41. Son or daughter leaving home (due to marriage, college, etc.)			-3	-2	-1	0	+1	+2	+3
42. Ending of formal schooling			-3	-2	-1	0	+1	+2	+3
43. Separation from spouse (due to work, travel, etc.)			-3	-2	-1	0	+1	+2	+3
44. Engagement			-3	-2	-1	0	+1	+2	+3
45. Breaking up with boyfriend/girlfriend			-3	-2	-1	0	+1	+2	+3
46. Leaving home for the first time			-3	-2	-1	0	+1	+2	+3
47. Reconciliation with boyfriend/girlfriend			-3	-2	-1	0	+1	+2	+3
<i>Other recent experiences which have had an impact on your life. List and rate.</i>									
48. _____			-3	-2	-1	0	+1	+2	+3
49. _____			-3	-2	-1	0	+1	+2	+3
50. _____			-3	-2	-1	0	+1	+2	+3

	0 to 6 mo	7 mo to 1 yr	extremely negative	moderately negative	somewhat negative	no impact	slightly positive	moderately positive	extremely positive
Section 2: Student Only									
51. Beginning a new school experience at a higher academic level (college, graduate school, professional school, etc.)			-3	-2	-1	0	+1	+2	+3
52. Changing to a new school at same academic level (undergraduate, graduate, etc.)			-3	-2	-1	0	+1	+2	+3
53. Academic probation			-3	-2	-1	0	+1	+2	+3
54. Being dismissed from dormitory or other residence			-3	-2	-1	0	+1	+2	+3
55. Failing an important exam			-3	-2	-1	0	+1	+2	+3
56. Changing a major			-3	-2	-1	0	+1	+2	+3
57. Failing a course			-3	-2	-1	0	+1	+2	+3
58. Dropping a course			-3	-2	-1	0	+1	+2	+3
59. Joining a fraternity/sorority			-3	-2	-1	0	+1	+2	+3
60. Financial problems concerning school (in danger of not having sufficient money to continue)			-3	-2	-1	0	+1	+2	+3

Received June 23, 1977 ■

Young Adult Schizophrenics: Prediction of Outcome and Antecedent Childhood Factors

James D. Roff
Eastern Michigan University

Raymond Knight
Brandeis University

A young adult sample of acute schizophrenics was followed through record sources into middle age. Antecedent childhood information was also obtained. Three aspects of schizophrenia (psychotic thinking, affectivity, and social competence) were assessed in terms of long-term stability and prediction of outcome criteria. Measures of psychotic thinking were found to lack both stability and predictive validity. In contrast, a combined measure of affectivity and social competence was stable over time and was significantly related to eventual outcome. Childhood factors were also related to adult outcome variables. Implications for research definitions of schizophrenia are discussed.

Strauss, Carpenter, and Bartko (1974) have emphasized three underlying processes or dimensions in schizophrenia: positive symptoms, negative symptoms, and disordered social relationships. Positive symptoms primarily involve psychotic thinking, more specifically, the presence of delusions and hallucinations. Negative symptoms refer to the absence of an attribute. Flat or blunted affect is the principal example of a negative symptom, that is, the absence of normal or appropriate affect. Disordered social relationships are related to social competence as measured by the Phillips scale (Phillips, 1953) and other comparable scales.

Strauss et al. (1974) have stated that positive symptoms appear to be prognostically neutral within a schizophrenic sample. Also, they consider that negative symptoms may be confounded with chronicity and possess reduced prognostic value in the absence of chronicity. Disordered social relationships have well-established prognostic validity for schizophrenics as reflected by studies that have used process-reactive or good-poor pre-morbid distinctions.

The present study used a sample of young adult schizophrenics, who in terms of length

of hospitalization and duration of psychosis were acute patients. This allowed the assessment of variables selected to measure psychotic thinking, affectivity, and social competence as prognostic indicators prior to the effects of extended hospitalization. Preschizophrenic information was available from child guidance clinic records. Details of the childhood data analysis have been reported by Roff, Knight, and Wertheim (1976). This prior study found four childhood factors that were then related to a global measure of long-term adult outcome. The present study, in contrast, has focused on factors found during the young adult period and investigated their prognostic significance in terms of subsequent clinical status during the middle adult period. In addition to a global measure of outcome, two factor scales provided more differentiated outcome criterion. The childhood factors, as well as the young adult factors, were assessed with a special interest in the possible differential prediction with regard to the two outcome factor scales.

Previous prognostic studies have indicated a number of significant variables (Nameche, Waring, & Ricks, 1964; Robins & Guze, 1970; Stephens, Astrup, & Mangrum, 1967; Vailant, 1964). These studies suggest that measures of social competence, both work and social adjustment, and measures of affectivity would be most appropriate given the data

Requests for reprints should be sent to James D. Roff, Department of Psychology, Eastern Michigan University, Ypsilanti, Michigan 48197.

available. It was also considered desirable to include measures of psychotic thinking, particularly those that reflect severity.

Related to the prognostic question is the problem of defining schizophrenia for research purposes. In this study, a hospital diagnosis of schizophrenia has been used. This is recognized as an overly inclusive definition but has the advantage that few probable schizophrenics would be initially excluded. Outcome ratings were made that amounted to a rediagnosis of each case based on follow-up information. In addition, the New Haven Schizophrenia Index (NHSI), which was developed as a checklist definition of schizophrenia for research use (Astrachan et al., 1972), was included. The NHSI is similar to other checklists for schizophrenia that are heavily weighted with items that reflect aspects of psychotic thinking. It is desirable for definitions of schizophrenia to involve aspects of the syndrome that possess long-term stability. The concept of process schizophrenia carries with it the notion that a stable underlying condition exists. This study focuses on the related problems of prediction of outcome and the search for definitional aspects of schizophrenia that have long-term stability.

Method

Subjects were 45 males who received a hospital diagnosis of schizophrenia while in the military service. A few cases were diagnosed as schizoid personality in service but were subsequently diagnosed as schizophrenic by the Veterans Administration (VA). Postservice follow-up was possible through the use of centralized VA files. Preservice information was independently obtained from child guidance clinic records. Three separate records sources provided data on the same individuals for childhood (M age = 10.9), young adult (M age = 21.7), and middle adult (M = 43.7) periods.

Postservice VA data were used to establish outcome groups to be used as predictive criteria for the service and childhood information. Each case was assigned a position on a 6-point scale, which reflected severity of impairment and certainty of process schizophrenia. Outcome ratings reflected the independent judgments of the two principal investigators. A consensus was reached for those cases with initial disagreement. Outcome 1 involved recovered cases ($n = 2$); Outcome 2 cases had minimal signs of schizophrenia at follow-up but had neurotic or acting-out character deficits ($n = 16$); Outcome 3

cases were paranoid with emotionally unstable personalities ($n = 6$); Outcome 4 cases had severe schizoid personalities but had minimal thought disorder ($n = 10$), and Outcomes 5 and 6 were considered process schizophrenics, with the former improved for substantial periods ($n = 9$) and the latter unimproved or deteriorated ($n = 4$). Independent clinical judgments that were used as the basis for the consensus outcome ratings had an interrater agreement of 76%. (For further information about the derivation of outcome ratings, see Roff et al., 1976.) Outcome ratings were significantly correlated with other indices of outcome—length of hospitalization, employment history, marital status, and level of VA compensation for psychiatric disability.

Service information was abstracted from military service records, with raters blind to both childhood and postservice information. Six variables were selected as the best available composite measures or global judgments of thought disorder, affectivity, and social competence. These variables included conceptual disorganization (Hautaluoma, 1971), a factor scale with item composition similar to a factor of the same name reported by Lorr, Klett, and McNair (1963); the NHSI (Astrachan et al., 1972); global clinical judgments of severity of thought disorder and severity of affect deficit; and ratings of service social adjustment and service work performance. The global clinical judgments were independently made by the two principal investigators on 5-point scales. The average of the two scores was used. Two research assistants independently used the factor scale and NHSI. The assistants also rated social adjustment and work performance on 3-point scales (good, fair, poor). Again, the average of two scores was used. Acceptable levels of interrater agreement were found for all of the measures, with the range of reliability coefficients from .70 to .97 with a median reliability of .82.

These six service variables were intercorrelated and then factor analyzed with significant factors rotated to an orthogonal, varimax solution. Scales were constructed by a summation of standard scores for relevant variables significantly associated with the obtained factors.

The same variables, but from postservice data, were used to construct additional parallel scales. In addition, the outcome rating, derived from postservice information, served as predictive criteria.

The childhood data had been previously analyzed with four factors reported (Roff et al., 1976). Factor scores were computed for factors of unsocialized aggressiveness, low IQ—poor school achievement, neuroticism, and schizoid syndrome.

The resulting set of variables was then analyzed in terms of longitudinal prediction. The set contained four childhood factor scores, two service scales, two postservice scales, and the outcome rating. Additionally, the stability of the two service scales was evaluated by means of their correlation with the two postservice scales.

Results

Service data were factor analyzed to determine a limited number of dimensions that would provide factor scales for use as variables in the prediction of postservice outcome measures. Table 1 shows the result of the factor analysis for the service data. The first factor was labeled *Psychotic Thinking*, with significant loadings for the variables of conceptual disorganization, NHSI, and global clinical judgments of severity of thought disorder. The second factor, labeled *Affect/Social Competence*, had significant loadings for global clinical judgments of affect deficit, service social adjustment, and service work performance. The scale scores, derived from the three variables loading significantly on each factor, retained the independence of the factor solution ($r = .17$). The first two factors accounted for 72% of the variance. The loading of the affect variable and the social competence variables on the same factor was particularly noteworthy along with the relative independence of the two factors when an oblique rotation was selected.

The factor structure for the postservice data was less differentiated. This was reflected by the fact that the two postservice scales, using similar variables, were significantly correlated ($r = .52$). The postservice scales were highly related to the outcome ratings (psychotic thinking and outcome = .77, affect/social competence and outcome = .73). At follow-up, thought disorder, affectivity, and social competence measures were all significantly related to each other. In contrast, dur-

Table 1
Orthogonal Factors for Service Variables

Variable	Factor	
	1	2
Work performance	-.05	.62*
Social adjustment	.15	.43*
Conceptual disorganization	.90*	.08
New Haven Schizophrenia Index	.77*	-.29
Clinical judgment		
Thought disorder	.65*	.36
Affect	.32	.93*

* $p < .01$.

Table 2
Relationship Between Service and Postservice Scales

Service	Postservice		
	1	2	Outcome
1. Psychotic thinking	.24	.11	.12
2. Affect/social competence	.44*	.51*	.65*

* $p < .01$.

ing the young adult period, the thought disorder measures were relatively independent of affectivity and social competence measures, whereas the later two measures were significantly correlated.

Table 2 shows the relationship of two service scales to the two postservice scales and to the outcome ratings. Clearly, it was the service scale measuring affect deficit and social competence that possessed both long-term stability ($r = .51$) and predictive validity ($r = .65$). In contrast, the Psychotic Thinking scale was neither stable nor predictive. Surprisingly, the Affect/Social Competence scale using service information was a better predictor of severity of psychotic thinking than the service Psychotic Thinking scale, which contained the same variables. Table 2 clearly demonstrates the unfavorable prognostic significance of the Affect/Social Competence factor during the service period, and the Psychotic Thinking factor failed to significantly predict any of the postservice measures.

Childhood factor scores were related to the outcome measures, with special interest directed to differential relationships with the two postservice factor scales. Correlations between the childhood factor scores and the postservice variables are presented in Table 3. Schizoid syndrome was significantly related to high scores on the Psychotic Thinking and Affect/Social Competence scales and to poor outcome. Unsocialized aggressiveness was related to more favorable outcome and more favorable scores on the Affect/Social Competence scale but was not significantly related to level of psychotic thinking. Low IQ - poor school achievement was related to unfavorable scores on the Affect/Social Competence scale. The pattern of correlations for the low

Table 3

Relationship Between Childhood Factor Scores and Postservice Scales

Childhood	Postservice		
	Psychotic thinking	Affect/social competence	Outcome
Unsocialized aggressiveness	-.12	-.35*	-.33*
Low IQ - poor school achievement	.06	.38*	.27
Neurotic symptoms	-.03	-.13	.04
Schizoid syndrome	.38*	.33*	.37*

* $p < .05$.

IQ - poor school achievement factor was all the more remarkable given the significant correlation between the two postservice factor scales. Level of neurotic symptoms did not predict any of the postservice variables. In general, the childhood variables predicted postservice better than service variables.

Classifying cases according to their combined score on the childhood schizoid syndrome and the service Affect/Social Competence scale produced the data in Table 4. Comparing the number of cases above and below the mean for each outcome group, Outcomes 1 and 2 combined were significantly different from Outcomes 5 and 6 combined. Outcomes 3 and 4 were more evenly distributed on their scores. Table 4 suggests that the poor outcome cases not only had high scores on the affect/social competence factor in service, but they also tended to have a poor pre-service history as measured by the childhood schizoid syndrome factor. This indicates a long-standing deficit that existed prior to extended hospitalization.

Although the number of cases was small, Outcome 3 cases were compared with Outcome 4 cases, using a point-biserial correlation with outcome dichotomized. Outcome 4 cases were strongly related to low IQ - poor school achievement childhood factor scores ($r_{pbis} = .68$). Outcome 3 cases were related to childhood unsocialized aggressiveness factor scores ($r_{pbis} = .37$). These relationships were consistent with the hostile-paranoid clinical picture for Outcome 3 and the severe schizoid or inadequate personalities characteristic of Outcome 4. Outcome 3 cases were not significantly different from Outcomes 1 and 2.

Outcomes 5 and 6 had higher scores on the childhood schizoid syndrome when compared with Outcome 4 cases but were not significantly different on the other three childhood variables. Outcomes 5 and 6 were consistently more disturbed than Outcome 4 cases both in service and postservice in terms of psychotic thinking and affect/social competence.

Table 5 summarizes the major findings that relate childhood and service factors to outcome. The neurotic and schizoid childhood factors and the Affect/Social Competence and Psychotic Thinking service factors were consistent in their relationships across all three outcome measures. In contrast, the unsocialized aggressiveness and low IQ - poor school achievement childhood factors were not significantly related to Psychotic Thinking scale scores but were significantly related, in opposite directions, to the Affect/Social Competence scale; and the former was significantly related to more favorable global outcome ratings while the latter approached significance in the unfavorable direction.

Table 4

Outcome by Schizoid Syndrome Plus Affect/Social Competence Scale

Outcome	Scale score	
	Above M	Below M
1, 2	2	14
3	2	4
4	4	6
5, 6	12	1

Note. $\chi^2(3) = 18.87, p < .001$.

Table 5
Childhood and Young Adult Factors in Relation to Outcome

Prognostic significance	Childhood factors	Young adult factors
Unfavorable	Schizoid syndrome Low IQ - poor school achievement ^a	Affect/social competence
Neutral Favorable	Neurotic symptoms Unsocialized aggressiveness ^b	Psychotic thinking

^a Unfavorably related to affect/social competence factor at outcome but neutral with respect to psychotic thinking and of borderline significance in relation to global outcome ratings.

^b Neutral with regard to psychotic thinking factor at follow-up.

Discussion

As with most schizophrenic samples, it is important to consider sample limitations. These include, in addition to sample size, a male sample, individuals with childhood problems (disturbed preschizophrenics), and the exclusion of cases rejected for military service. On the other hand, comparisons have been restricted to within-sample differences for a group that has had similar experiences in terms of child clinic contact, military service hospitalization, and the effects of being labeled schizophrenic. Results were consistent with the claim of Strauss et al. (1974) that positive symptoms are prognostically neutral within a schizophrenic sample. Affect deficit, primarily flat affect, measured the principal negative symptom. In this study, flat or blunted affect was not a product of extended hospitalization. Its appearance early in a subject's career was an unfavorable prognostic sign. Disordered social relationships were related to the clinical judgments of affect deficit. It is plausible that affect deficit arises as a result of a history of impaired social relationships. Flat affect may be a function of chronicity measured not in duration from the onset of psychotic thinking or positive symptoms but from the onset of extreme social maladjustment.

The childhood schizoid factor contained a measure of peer adjustment during childhood and a schizoid scale that included symptoms of apathy, flat affect, and seclusiveness. In other words, the schizoid syndrome included the childhood variables most similar to those in the Affect/Social Competence scale for the adult periods. Unsocalized aggressiveness in

childhood was most common in Outcomes 1, 2, and 3. The primary difference between these groups was the significantly higher level of psychotic thinking at follow-up for the Outcome 3 cases. Otherwise, Outcome 3 cases had a developmental background very similar to the more favorable Outcomes (1 and 2). Outcome 4 cases presented a contrast to Outcome 3 at follow-up with lower levels of psychotic thinking but had higher scores on the Affect/Social Competence scale. During childhood, the Outcome 4 cases were less aggressive than Outcome 3 cases and more inadequate as reflected by lower IQ and poorer school achievement. These results suggest that Outcome 3 cases might be considered poor outcome members of a larger group with generally favorable outcome, whereas Outcome 4 cases might have more favorable outcomes from a group with generally poor outcome.

Attempts to define schizophrenia that are heavily weighted with psychotic thinking such as the NHSI appear less promising than alternative measures of affectivity and social competence. Social competence has been adequately assessed by existing scales. A recent scale developed by Chapman, Chapman, and Raulin (1976) to measure anhedonia may provide a more convenient measure of affectivity. This study suggests that the combination of a social competence measure with one of affectivity may assess an important definitional aspect of schizophrenia with both long-term stability and prognostic relevance. It should be noted that measures of psychotic thinking during the postservice period were highly related to measures of affectivity, social competence, and outcome. It was during

the service period that psychotic thinking was a poor substitute for the other measures. Of course, it is the early stage in which problems of definition and prediction are most challenging.

The results of this study do indicate the stability of some behaviors of schizophrenic subjects over extended periods of time. It appears that adequate assessment of these stable aspects of schizophrenia should be incorporated into research definitions of schizophrenia.

References

- Astrachan, B. M., et al. A checklist for the diagnosis of schizophrenia. *British Journal of Psychiatry*, 1972, 121, 529-539.
- Chapman, L. J., Chapman, J. P., & Raulin, M. L. Scales for physical and social anhedonia. *Journal of Abnormal Psychology*, 1976, 85, 374-382.
- Hautaluoma, J. Symptoms, antecedents, and outcomes of psychosis: A cluster analytic study. *Journal of Consulting and Clinical Psychology*, 1971, 37, 332-344.
- Lorr, M., Klett, C. J., & McNair, D. *Syndromes of psychosis*. New York: Macmillan, 1963.
- Nameche, G., Waring, M., & Ricks, D. Early indicators of outcome in schizophrenia. *Journal of Nervous and Mental Disease*, 1964, 139, 232-240.
- Phillips, L. Case history data and prognosis in schizophrenia. *Journal of Nervous Mental Disease*, 1953, 117, 515-525.
- Robins, E., & Guze, S. Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American Journal of Psychiatry*, 1970, 126, 983-987.
- Roff, J. D., Knight, R., & Wertheim, E. A factor analytic study of childhood symptoms antecedent to schizophrenia. *Journal of Abnormal Psychology*, 1976, 85, 543-549.
- Stephens, J. M., Astrup, C., & Mangrum, J. C. Prognosis in schizophrenia. *Archives of General Psychiatry*, 1967, 16, 693-698.
- Strauss, J. S., Carpenter, W. T., & Bartko, J. J. An approach to the diagnosis and understanding of schizophrenia: Part III. Speculations on the processes that underlie schizophrenic symptoms and signs. *Schizophrenia Bulletin*, 1974, 11, 61-69.
- Vaillant, G. E. Prospective prediction of schizophrenic remission. *Archives of General Psychiatry*, 1964, 11, 509-518.

Received June 24, 1977 ■

Blood Alcohol Level Discrimination by Alcoholics: The Role of Internal and External Cues

David Lansky, Peter E. Nathan, and David M. Lawson
Rutgers—The State University

Two groups of four chronic alcoholic subjects lived in Rutgers' Alcohol Behavior Research Laboratory for separate 2-week periods. During that time, subjects were taught to attend either to internal or to external cues to blood alcohol level (BAL). During a single training session, subjects received feedback on actual BAL following each of their BAL estimates. During pretraining and posttraining sessions, assessments of BAL estimation accuracy were obtained in the absence of feedback. Prior to training, both groups of alcoholics were equally inaccurate in estimating BAL. During training, when accurate BAL feedback was provided, estimation accuracy increased significantly for both groups. Once feedback of actual BAL was removed during the posttraining test session, however, only externally trained subjects maintained the ability to estimate BAL accurately. It was concluded that unlike the nonalcoholic subjects studied by Huber, Karlin, and Nathan, the alcoholic subjects of this research did not learn to discriminate BAL on the basis of internal feelings and sensations nearly as adequately as they did when they referred to external cues. These findings have important implications for the clinical application of BAL discrimination training.

In recent years the widespread conviction that abstinence constitutes the only legitimate treatment goal for alcoholism has come under increasing scrutiny. Acceptance of abstinence-oriented treatment goals for all alcoholics has been challenged by findings that some alcoholics can acquire and maintain patterns of moderate social drinking without accompanying "loss of control" over intake (Armor, Polich, & Stambul, 1976; Davies, 1962; Pattison, 1968; Popham & Schmidt, 1976) and by the apparent success of alcoholism treatment programs with controlled drinking as an explicit treatment goal (Lovibond & Caddy, 1970; Pomerleau, Pertschuk, & Stinnet, 1976; Sobell & Sobell, 1973, 1976).

Several studies on the utility of controlled drinking-oriented treatment for alcoholism incorporate blood alcohol level (BAL) discrimination training as a component. The most frequently used BAL discrimination training method involves the delivery of accurate feedback on BAL under training conditions designed to sensitize the individual to the affective and physiological ("internal") concomitants of different BALs. Caddy and Lovibond (1976), Vogler, Compton, and Weissbach (1975), and Wilson and Rosen (1975) have concluded that this way to train BAL discrimination gives the alcoholic the ability to monitor level of intoxication and then to use this newfound skill to maintain more moderate BALs.

A review of the basic research findings relevant to this treatment approach, however, failed to confirm that alcoholics can in fact acquire and maintain the ability to discriminate BAL on the basis of internal cue training alone. For example, the positive results of studies by Lovibond and Caddy (1970), Vogler et al. (1975), and Paredes, Jones, and Gregory (1974) are all difficult to interpret

This study was supported by National Institute on Alcohol Abuse and Alcoholism Grant AA00259-07 to Peter E. Nathan. We thank John Miller, Department of Statistics, Rutgers—The State University, and the staff of the Alcohol Behavior Research Laboratory for their essential help in this project.

Requests for reprints should be sent to Peter E. Nathan, Alcohol Behavior Research Laboratory, Rutgers—The State University, New Brunswick, New Jersey 08903.

Table 1

Means for Demographic Characteristics and Drinking Histories of Subjects in the Internal and External Training Groups

Variable	Training		<i>t</i>
	Internal	External	
Age (years)	36.25	35.75	.141
No. years of education	11.75	10.75	.880
No. alcohol-related hospitalizations or treatment programs	1.50	1.75	.205
No. years of problem drinking	15.00	9.50	1.166

Note. $n = 4$. All *t* tests are nondirectional tests for independent samples. $df = 6$.

unequivocally, since BAL discrimination accuracy was not assessed in any of the three studies prior to training or after it was terminated. As a result, it is impossible to know whether the alcoholic subjects of these investigations actually did (a) improve their capacities to estimate BAL accurately and (b) maintain this accuracy when external cues to intoxication (e.g., drink strength, veridical BAL feedback, etc.) were removed.

An investigation designed to study the impact of some of these factors on BAL estimation was recently published by Silverstein, Nathan, and Taylor (1974). Alcoholics were asked to provide BAL estimates on the basis of internal feelings and sensations both when accurate feedback on BAL was provided (during internal cue training itself) and when it was not (during pretraining and posttraining sessions). Although discrimination accuracy improved from the initial baseline period to the subsequent period of training, removal of feedback during a final posttraining phase of the study resulted in a return to pretraining levels of estimation accuracy. In essence, then, this initial pre-post comparison of BAL discrimination by alcoholics did not support the view that alcoholics can learn to discriminate BAL on the basis of internal cues alone.

In contrast, the results of BAL discrimination studies of nonalcoholic social drinkers (e.g., Bois & Vogel-Sprott, 1974; Huber, Karlin, & Nathan, 1976) suggest that this subject population can learn to discriminate BAL on the basis of subjective feelings and sensations. In fact, Huber et al. (1976) found that social drinkers can learn to discriminate BAL

equally well when trained to attend either to internal cues (feelings and sensations) or to external ones (BAL-dose relationships).

To date, however, no study of alcoholic subjects has rigorously explored the efficacy of BAL discrimination training focused on external cues nor has external cue training yet been compared directly to internally focused training. Nonetheless, as noted above, clinical studies of BAL discrimination continue to report attempts to train alcoholics to discriminate BAL via internal cues in the absence of empirical evidence that they can in fact do so. This critical research lacuna emphasizes the importance of a direct comparison of these two BAL discrimination training methods with alcoholics.

The present study was designed to effect just such a comparison. To this end, alcoholic subjects were selected for an essential replication of the Huber et al. (1976) study, which compared the efficacy of external cue training and internal cue training with nonalcoholic social drinkers. It was hypothesized that unlike the nonalcoholic subjects studied by Huber and his colleagues, the alcoholic subjects in this study would be less well able to acquire accurate BAL discrimination on the basis of internal than external training procedures.

Method

Subjects

Subjects were recruited via advertisements placed in regional newspapers. They were offered \$60 per

week to live in the Alcohol Behavior Research Laboratory (described in Nathan, Goldman, Lisman, & Taylor, 1972) and participate in the study.

Four men were selected for each of two training groups. All subjects met the following criteria: (a) more than 2 years of heavy problem drinking; (b) evidence of physical dependence on and tolerance to ethanol; (c) good physical health, with no signs of liver or kidney damage; (d) no evidence of psychosis or chronic brain syndrome; (e) not dependent on prescription or street drugs at the time of the research. A complete physical examination, with appropriate laboratory tests, was given to each subject. Demography and drinking history of the two subject samples is summarized in Table 1. Nondirectional *t* tests for independent samples revealed no significant differences between the two groups on any of these variables.

Apparatus

BALs were measured from breath samples by the Gas Chromatograph Intoximeter, Mark IV (Intoximeters Inc., St. Louis, Missouri).

Procedure

Each subject participated in a total of three sessions. From midnight of the day preceding each of the three sessions, subjects were deprived of all food and beverage. The morning of the session, each subject received one cup of decaffeinated coffee; lunch was served at the end of the session.

Session 1 (pretraining). Session 1 was a baseline session, designed to assess subjects' pretraining BAL estimation accuracy. All subjects consumed a total of six drinks containing 80-proof vodka and tomato juice at 35-minute intervals. Each drink contained a mixture of either $\frac{1}{2}$, 1, or $1\frac{1}{2}$ ounces of vodka, plus enough tomato juice to make a total of 6 ounces of liquid. In this and the other two sessions, these three dosage levels were randomly distributed, with the $\frac{1}{2}$ -ounce dose given once, the 1-ounce dose given twice, and the $1\frac{1}{2}$ -ounce dose given three times in each series of drinks. Accordingly, total consumption of liquid was 36 ounces in a 3-hour period, 7 ounces of which were vodka. Following each drink, a subject completed a postdrink questionnaire, designed to assess his ability to discriminate the amount of alcohol contained in the drink.

Four BAL estimates were made during the course of this session. Neither BAL feedback nor training in BAL estimation were provided during the session. Subjects were required to make their estimates on a scale of 0-150, with 0 representing "cold sober" and 150 representing "very high, about as high as you've ever been." Estimates were scheduled 25 minutes after the second and fifth drinks and 1 hour and $1\frac{1}{2}$ hours after the sixth drink.

Subjects in all sessions spent the time between drinks and BAL estimates in individual bedrooms.

Session 2 (training). Sessions 1 and 2 were separated by a day, during which alcohol was not available.

Drink doses were randomized as they had been during Session 1. Ingestion of drinks took place as it did during the first session, except that subjects were now required to gargle with an anesthetic mouthwash prior to each drink; the gargling served briefly to anesthetize the subject's mouth and upper throat, thereby interfering with his ability to use taste cues to discriminate the amount of alcohol in the drink. As in Session 1, a postdrink questionnaire was completed immediately after a subject had finished each drink.

Depending on the group to which he was assigned, a subject received training in BAL estimation that focused either on external or on internal cues.

External training. Each subject in this group received a programmed learning booklet explaining BAL-dose relationships. The booklet was prepared for each subject individually so that the approximate absorption rate of alcohol that he used was based on his weight and resultant BALs as observed in Session 1. All subjects were taught that they metabolize alcohol at the rate of about 10 points per hour on the 0-150 scale and that an ounce of 80-proof vodka produces a rise of 8, 10, or 12 points, depending on weight and metabolism rate. All subjects were allotted sufficient time (2 hours) to complete their booklets. The results of written tests, administered following completion of the booklet, indicated that all subjects had mastered the material. The training sequence for the remainder of this session was as follows: (a) at bar, gargling and drinks, postdrink questionnaire; (b) in room, 20-minute wait; (c) at Intoximeter, told alcohol content of immediately preceding drink, BAL estimate, BAL feedback; (d) drink sequence repeated.

Feedback from the Intoximeter was converted to the subject's scale of 0-150 (0 = 0 mg%; 150 = 300 mg%).

Internal training. Internal training began with each subject listening to a standard relaxation tape designed to increase awareness of important muscle groups and of the sensations that arise from them during states of tension and relaxation. The tape was played once for 15 minutes. Subjects were asked to sit quietly, focus on, then "tune-in" to the bodily sensations that they were experiencing. To help subjects become further aware of internal states, they were also asked to complete modified versions of the Body Sensation Checklist (Bois & Vogel-Sprott, 1974) and the Mood Adjective Check List (McNair & Lorr, 1964) at this time. Prior to each BAL estimate, the subjects in this group again went through all procedures described above except for listening to the relaxation tape. Accordingly, the experimental sequence for this group was as follows: (a) At bar, gargling and drinks, postdrink questionnaire; (b) in room, 5-minute wait, Mood Adjective Check List, tune-in instructions, Body Sensation Check List; (c) at Intoximeter, BAL estimate, BAL feedback, tune-in instructions; and (d) drink sequence repeated. As

with the external training group, BAL feedback was converted to the 0-150 scale.

The drink-estimate sequence was repeated six times during this session, with two additional estimates programmed 1 and 1½ hours following the last drink.

Session 3 (testing). This session followed Session 2 by a day, during which alcohol was unavailable. Procedurally, it was identical to the first session with two exceptions. First, all subjects were required to continue gargling with the anesthetic mouthwash prior to each drink. Second, subjects in both groups were told the actual alcoholic content of their drinks after completing the postdrink questionnaire. This was done to standardize testing conditions and to ensure that Session 3 provided a fair test of the efficacy of both training procedures. Thus, by making dosage information available, externally trained subjects were provided the minimal information necessary to appropriately utilize their training, whereas internally trained subjects were provided more than sufficient information to fully utilize internal training methods. As a result of this testing procedure, outcome was biased in favor of the internal training method. After each BAL estimate, each subject was required to fill out a postdecision questionnaire, indicating how he had arrived at his particular estimate of BAL. Finally, all subjects were told that they would receive a \$3 bonus for each estimate given this session that was within 5 points of their actual BAL.

Data Analysis

BAL estimation accuracy. BAL estimation accuracy was defined as the absolute difference between actual and estimated BAL. In order to have comparable data points across sessions, only those four BAL estimates in Session 2 (2, 5, 7, and 8) that corresponded in time to those in Sessions 1 and 3 were included in the data analysis. Sessional changes in error scores were analyzed by means of a three-factor analysis of variance (Edwards, 1972), with sessions, groups, and trials (i.e., estimates) as main effects. Within-session data were analyzed by means of two-factor analyses of variance, with groups and trials as main effects. Data analyses included computation of partial correlations (Ferguson, 1971) between actual BAL and BAL estimates, with the effect of between-subject variability partialled out. These analyses were performed for all subjects within a group on a session-by-session basis. Third session mean error scores for the two groups were also compared by means of analysis of covariance (Edwards, 1972), with mean Session 1 error scores as the covariate.

Discriminability of drink strength. To determine whether the three alcohol doses were discriminable from each other, the drink estimates of both training groups were combined for analysis purposes, yielding estimates for eight ½-ounce drinks, 16 1-ounce drinks, and 24 1½-ounce drinks in each session. These within-session data were then subjected to one-factor analyses of variance for unequal *ns* (Edwards, 1972),

with the three dosage levels analyzed to determine the single main effect of dose. Relative drink discrimination accuracy by each of the two training groups was analyzed by comparing drink estimation error scores (defined as the absolute value of the difference between a subject's estimate of the alcoholic content of a drink and actual drink dosage) at all three dosage levels. To this end, within-session error scores were subjected to two factor analyses of variance, with groups and trials as main effects.

Results

BAL Estimation Accuracy

Raw BAL estimation error scores could not be subjected directly to analysis of variance, since an F_{\max} test (Winer, 1962) revealed significant heterogeneity of variance across sessions ($F_{\max} = 118.62$, $p < .01$). A logarithmic transformation (\log actual error score + 1.0) was required to reduce the heterogeneity of variance.

Figure 1 graphs the mean transformed error score of each group in each session. As this

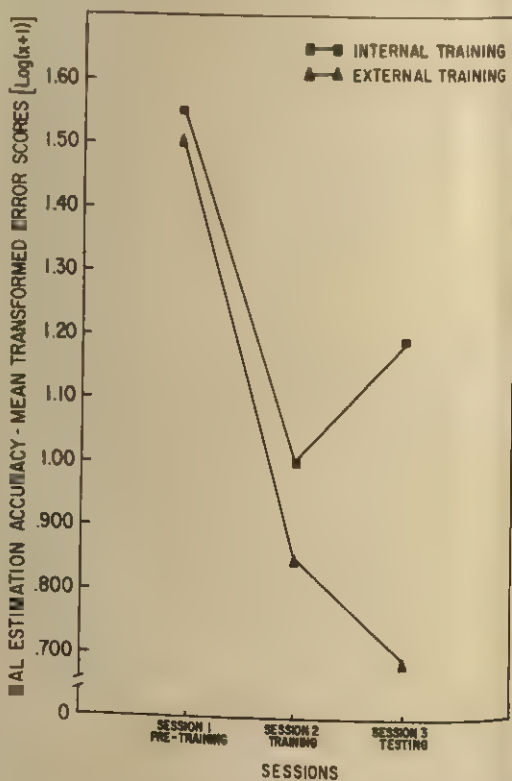


Figure 1. Mean transformed blood alcohol level (BAL) estimation error scores of the alcoholics by session and group.

figure shows, both groups markedly improved estimation accuracy with training. Mean transformed error scores, averaged over groups, were 1.5202 for Session 1, .9548 for Session 2, and .9374 for Session 3. An analysis of variance of these data revealed a significant sessions effect, $F(2, 12) = 5.49$, $p < .02$. Duncan's multiple-range test for differences in session means (Edwards, 1972) revealed significant differences between the mean error score in Session 1 and those in Sessions 2 and 3. Mean error scores in Sessions 2 and 3 did not differ from each other. These results confirm that BAL estimation accuracy improved significantly for both groups with training.

Session 1 (pretraining). No significant effects were revealed by the analysis of variance applied to Session 1 data. Mean transformed error score for internally trained subjects was 1.5368; for externally trained subjects, it was 1.5036. The relationship between BAL estimates and actual BAL was also examined by computing partial correlation coefficients (see Table 2). Although the correlation for the external-training group (.47) was greater than that for the internal-training group (.23), neither of these correlations approached significance. If the degree of correlation between actual and estimated BAL is considered a measure of the ability to accurately monitor changes in actual BAL, then this analysis suggests that neither group of alcoholics was able to monitor these changes to a significant degree during this pretraining session.

Session 2 (training). As noted above and shown in Figure 1, the BAL estimation accuracy of both groups improved markedly in Session 2. A within-session analysis of variance revealed no significant effects for this session; the difference between the mean transformed error score for internally trained subjects of 1.0545 and the score of .8550 for externally trained subjects was not significant. In addition to these lowered error scores, both training groups were more successful at monitoring actual BAL, as evidenced by increased correlations between BAL estimates and actual BAL.

Table 2

Partial Correlations Between Actual and Estimated Blood Alcohol Level for Each Group on a Session by Session Basis

Session	Internal	External
1	.23	.47
2	.66**	.52*
3	-.06	.91***

* $p < .07$.

** $p < .02$.

*** $p < .001$.

Session 3 (testing). An analysis of variance applied to Session 3 data revealed a significant group effect, $F(1, 6) = 7.69$, $p < .035$. No other effects reached significance. The mean transformed error score for the internal-training group was 1.1860; it was .6887 for the external-training group. Mean Session 3 error scores were also analyzed with mean Session 1 error scores as the covariate. This analysis also revealed strong group differences, $F(1, 5) = 6.41$, $p < .053$.

Group differences in Session 3 estimation accuracy appear even more striking when one considers correlations between actual and estimated BALs (see Table 2). Although a high correlation between these scores was attained by internally trained subjects during Session 2, a similar analysis of Session 3 data failed to reveal a similarly high significant correlation for these subjects. By contrast, the correlation attained by the externally trained group in Session 3 was substantially higher than that observed during Session 2.

Discriminability of Drink Strength

An analysis of variance applied to within-session drink estimation data revealed significant differences between estimates given at each dosage level during all three sessions, $F(2, 45) = 14.53$ (Session 1); 4.02 (Session 2); 15.41 (Session 3), all $ps < .03$. Duncan's multiple-range test indicates that estimates at the $\frac{1}{2}$ -ounce dose were less than those at the 1-ounce dose, which, in turn, were less than those at the $1\frac{1}{2}$ -ounce dose (all $ps < .05$). These data show, then, that even when subjects gargled with an anesthetic mouthwash

prior to consuming each drink during Sessions 2 and 3, they could discriminate among drinks of varying doses. When drink estimation error scores of the two groups were compared, there were no group differences in drink estimation accuracy during Sessions 1 and 2. During Session 3, however, internally trained subjects tended to be more accurate (they had lower error scores) than externally trained subjects, $F(1, 6) = 5.73, p < .053$.

Self-Report Measures

Following each Session 3 BAL estimate, subjects were required to describe the means by which that estimate had been made. Externally trained subjects reported having used only the training method that they had been taught during Session 2. Internally trained subjects also reported their estimates, depending on their respective training method; two of these subjects noted, in addition, that they had attended to the number of drinks consumed as a cue to their estimates. Subjects were also asked to describe any difficulties that they had experienced with their respective training modes. No externally trained subject reported any difficulty in this respect. By contrast, internally trained subjects reported that the gargling had detracted from the accuracy of their BAL estimates by masking the taste of their drinks.

Discussion

For the chronic alcoholic subjects who took part in this study, there appears to be little doubt that training in BAL discrimination via external cue training was more effective than training in internal cues. Although all eight alcoholic subjects demonstrated comparable pretraining levels of estimation accuracy and comparable levels of accuracy during a subsequent training session, externally trained subjects were significantly more accurate in their estimates after training ended and feedback was withdrawn.

The relative superiority of the external training mode for these subjects is also reflected by the results of correlational analyses. As noted above, we presume that the de-

gree of correlation between actual and estimated BALs is a direct reflection of the accuracy with which subjects monitored changes in BAL. In this regard, although externally trained subjects were more accurate in the ability to monitor changes in BAL before training began, neither group of subjects was able to do so particularly well at this point. In Session 2, however, when feedback on accuracy in the context of training was provided, internally trained subjects markedly improved their monitoring performance; in fact, their accuracy at this point was marginally beyond that attained by externally trained subjects whose Session 2 accuracy was not substantially improved over that of Session 1. However, once training ended and feedback was removed (during Session 3), the externally trained group substantially improved the accuracy of its BAL monitoring (to .91), whereas the correlation between actual and estimated BAL for the internally trained group returned to its Session 1 level of essentially zero.

The pattern of correlations observed for the externally trained group suggests that BAL feedback alone did not enhance monitoring performance; rather, practice with the external training method itself appears to have been largely responsible for the development and maintenance of highly accurate monitoring. Conversely, as soon as internal training was begun, accurate tracking was observed immediately, but it was maintained for only as long as veridical feedback was available to the internally trained subjects. Thus, monitoring of BAL on the basis of external cues appears to have developed more slowly but to have been more enduring, in large part because it is less dependent on feedback than monitoring by subjects instructed to attend to internal cues.

Several methodological problems may affect the generality of these findings. First, internally trained subjects were provided drink dosage information during Session 3 but not during Session 2; this difference in training and testing conditions may have interfered with the internal discrimination techniques that these subjects were taught. This degree of impact seems unlikely, however, since self-

report questionnaire data indicated that internally trained subjects attended primarily to their feelings and sensations in formulating Session 3 BAL estimates and that external cues, when available, facilitated this discrimination process. In addition, since subjects were able to discriminate among drinks of varying doses in all three sessions, it is unlikely that this procedural variant provided information to which internally trained subjects did not already have access.

A second methodological question that can be raised concerns the specificity of the training information provided to the two groups of subjects. In this regard, externally trained subjects were provided objective information concerning relationships between distinct external cues (i.e., alcohol doses) and BALs. By contrast, internally trained subjects were required to learn their discrimination task using cues that were less distinct and objectifiable. Some authors (e.g., Lovibond & Caddy, 1970) have attempted to circumvent this problem by providing internally trained subjects with some objective information about the relationship between internal cues and blood alcohol levels. (e.g., "At a BAL of .05, you will begin to feel a little unsteady.") This procedure was not followed in the present study because it was felt that if internal cues were linked to different BALs for different subjects, as suggested by Huber et al. (1976), the provision of standardized information might have confused individual subjects by providing them with information that might not have been relevant in their particular cases.

The small number of subjects in our sample limits generalizations that can be drawn from our data. However, these data substantially support preliminary findings of the only other study of BAL discrimination by alcoholics that programmed pretraining and post-training tests of discrimination accuracy (Silverstein et al., 1974). It would appear to be a reliable finding, therefore, that alcoholics are relatively unable to discriminate BAL on the basis of internal feelings and sensations. These findings would bear considerable diagnostic significance if it could also be shown that nonalcoholics have less difficulty discrimi-

nating BAL on the basis of internal cues. For example, Huber et al. (1976), Ludwig and Wikler (1974), Pattison (1976), and Silverstein et al. (1974) have all observed that insensitivity to the visceral events that accompany alcohol intake, as reflected by failure to discriminate BAL on the basis of internal cues, may be functionally related to an alcoholic's inability to control his or her drinking. As already noted, Huber et al. (1976) and Bois and Vogel-Sprott (1974) have reported that nonalcoholics can discriminate BAL on the basis of internal cues. A recent study of nonalcoholics by Maisto and Adesso (1977), however, failed to support these earlier findings. In their study, nonalcoholic social drinkers accurately discriminated BAL only when they knew that they were consuming alcohol; when misled into thinking that they were drinking tonic, the discrimination accuracy of these subjects diminished substantially. It is not clear, therefore, whether the discrimination accuracy of Maisto and Adesso's social drinkers—as well as that of subjects studied by Huber et al. (1976) and Bois and Vogel-Sprott (1974)—was due to discriminated sensitivity to different BALs or to other factors such as subjects' knowledge of BAL dose-response curves and/or their acquiescence to demand characteristics of the experimental paradigm. Given these inconsistencies in data bearing on BAL discrimination skills of social drinkers, the diagnostic significance of alcoholics' apparent inability to discriminate BAL remains equivocal, though their behavioral difference from nonalcoholics in our BAL training paradigm remains real.

The findings reported here bear substantial clinical significance. In the first place, despite elaborate "internal" training procedures developed by other investigators to induce discriminated sensitivity to a range of BALs, our data suggest that a simpler, more effective means of accomplishing the same end is to train alcoholics to attend to the external cues that accompany alcohol intake. More generally, alcoholics' apparent insensitivity to the internal cues that accompany alcohol intake suggests that therapy designed to enhance consummatory self-regulation would be more

properly approached via interventions that focus on external rather than internal sources of control. This conclusion, in turn, suggests that behavioral approaches to alcoholism treatment that focus on environmental determinants of drinking behavior might be more effective with problem drinkers than more traditional insight-oriented therapies.

References

- Armor, D. J., Polich, J. M., & Stambul, H. B. *Alcoholism and treatment*. Santa Monica, Calif.: Rand, 1976.
- Bois, C., & Vogel-Sprott, M. Discrimination of low blood alcohol levels and self-titration skills in social drinkers. *Quarterly Journal of Studies on Alcohol*, 1974, 35, 86-97.
- Caddy, G. R., & Lovibond, S. H. Self-regulation and discriminated aversive conditioning in the modification of alcoholics' drinking behavior. *Behavior Therapy*, 1976, 7, 223-230.
- Davies, D. L. Normal drinking by recovered alcohol addicts. *Quarterly Journal of Studies on Alcohol*, 1962, 23, 94-104.
- Edwards, A. E. *Experimental design in psychological research*. New York: Holt, Rinehart & Winston, 1972.
- Ferguson, G. A. *Statistical analysis in psychology and education*. New York: McGraw-Hill, 1971.
- Huber, H., Karlin, R., & Nathan, P. E. Blood alcohol level discrimination by non-alcoholics: The role of internal and external cues. *Journal of Studies on Alcohol*, 1976, 37, 547-549.
- Lovibond, S. H., & Caddy, G. R. Discriminated aversive control in the moderation of alcoholics' drinking behavior. *Behavior Therapy*, 1970, 1, 437-444.
- Ludwig, A. M., & Wikler, H. "Craving" and relapse to drink. *Quarterly Journal of Studies on Alcohol*, 1974, 35, 108-130.
- Maisto, S. A., & Adesso, V. J. The effect of instructions and feedback on BAL discrimination training in nonalcoholic social drinkers. *Journal of Consulting and Clinical Psychology*, 1977, 45, 625-636.
- McNair, D. M., & Lorr, M. An analysis of mood in neurotics. *Journal of Abnormal and Social Psychology*, 1964, 69, 620-627.
- Nathan, P. E., Goldman, M. S., Lisman, S. A., & Taylor, H. A. Alcohol and alcoholics: A behavioral approach. *Transactions of the New York Academy of Sciences*, 1972, 34, 602-627.
- Paredes, A., Jones, B., & Gregory, D. An exercise to assist alcoholics to maintain prescribed levels of intoxication. *Alcohol Technical Reports*, 1974, 2, 24-36.
- Pattison, E. M. A critique of abstinence criteria in the treatment of alcoholism. *International Journal of Social Psychiatry*, 1968, 14, 268-276.
- Pattison, E. M. Nonabstinent treatment goals in the treatment of alcoholics. In R. J. Gibbins et al. (Eds.), *Research advances in alcohol and drug problems*. New York: Wiley, 1976.
- Pomerleau, O., Pertschuk, M., & Stinnet, J. A critical examination of some current assumptions in the treatment of alcoholism. *Journal of Studies on Alcohol*, 1976, 37, 849-867.
- Popham, R. E., & Schmidt, W. Some factors affecting the likelihood of moderate drinking by treated alcoholics. *Journal of Studies on Alcohol*, 1976, 37, 868-882.
- Silverstein, S. J., Nathan, P. E., & Taylor, H. A. Blood alcohol level estimation and controlled drinking by chronic alcoholics. *Behavior Therapy*, 1974, 5, 1-15.
- Sobell, M. B., & Sobell, L. C. Alcoholics treated by individualized behavior therapy: One-year treatment outcome. *Behaviour Research and Therapy*, 1973, 11, 599-618.
- Sobell, M. B., & Sobell, L. C. Second-year treatment outcome of alcoholics treated by individualized behavior therapy: Results. *Behaviour Research and Therapy*, 1976, 14, 195-215.
- Vogler, R. E., Compton, J. V., & Weissbach, T. A. Integrated behavior change techniques for alcoholics. *Journal of Consulting and Clinical Psychology*, 1975, 43, 233-243.
- Wilson, G. T., & Rosen, R. T. Training controlled drinking in an alcoholic through a multi-faceted behavioral treatment program: A case study. In J. D. Krumboltz & C. E. Thoresen (Eds.), *Counseling methods*. New York: Holt, Rinehart & Winston, 1975.
- Winer, A. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Received July 18, 1977 ■

Cuban Value Structure: Treatment Implications

José Szapocznik, Mercedes Arca Scopetta, Maria de los Angeles Aranalde
Department of Psychiatry, University of Miami

William Kurtines
Florida International University

This article discusses the relationship between cultural variables and psychosocial treatment. It is assumed that in order for psychosocial treatment to be acceptable and effective with a client population, it must be sensitive to the cultural characteristics of that population. The paradigm of planning therapy according to the cultural characteristics of a population is illustrated for Cuban immigrant adolescents. To investigate cultural variables, a Value Orientations Scale was developed based on the work of Kluckhohn and Strodtbeck using 325 subjects. Four factorially derived subscales were obtained. When 208 additional Cuban immigrant and Anglo-American adolescents were compared along the Value Orientations Scale, the Cubans tended to prefer lineality, subjugation to nature, present time, and not to endorse idealized humanistic values, whereas the Americans tended to prefer individuality, mastery over nature, future time, and to endorse idealized humanistic values. The implications of these value differences for the delivery of mental health treatment are discussed.

Cross-cultural conditions have seldom been investigated as variables of individual differences related to the appropriateness of different mental health treatment models. Recently, however, cultural variables have been considered as constituting relevant personal and situational characteristics that require specific culturally sensitive treatment approaches (Weidman, 1975).

In general, the issue of matching clients and treatment techniques to enhance the likelihood of obtaining desired outcomes has received extensive discussion and widespread endorsement in psychotherapy. Paul (1969), for example, has argued that psychotherapy outcome research should be directed toward ascertaining which treatment by whom is most effective for a person with specific characteristics and problems in a particular set of cir-

cumstances. This paradigm for treatment outcome research has received strong support from Bergin (1971), and Kiesler (1969, 1971) among others.

During the past 18 years, almost 700,000 Cubans have migrated to the United States. Approximately 500,000 have settled in the Greater Miami area, comprising about 90% of the local Latin population. Establishing mental health and drug abuse treatment services for the Cuban community have presented serious problems for the providers of these services, because the Cubans did not seek treatment from the established Anglo-American-oriented programs (cf. Ladner, Page, & Lee, 1975). These patterns of health care utilization are consistent with those observed in other Latin groups who, in general, underuse Anglo-American-oriented mental health services (Padilla & Ruiz, 1973). Concomitant with the immigrant status of the Cubans, high levels of behavioral disorders were expected to occur as had been found with other immigrant groups (Al-Issa, 1970; Berry & Annis, 1974; Mezey, 1960). It was urgent, therefore, to develop therapeutic models fea-

This work was supported by National Institute on Drug Abuse Grant 5 H81 DA 01 696. The authors are grateful to Robert Allen for his editorial assistance.

Requests for reprints should be sent to José Szapocznik, Spanish Family Guidance Clinic, 2121 Southwest 27 Avenue, Miami, Florida 33145.

sible for attracting and maintaining these Cubans in treatment.

The Problem

The present study is based on the assumption that in order to develop therapeutic models that will effectively attract and maintain clients in therapy, their cultural background must be understood. Specifically, it is postulated that cultural variables constitute an important set of client characteristics that need to be taken into consideration for developing valid statements about the relationship between adolescent Cuban clients in treatment and the appropriateness of treatment models (cf. Kiesler, 1971). It is further hypothesized that an understanding of the cultural differences between Cuban immigrant and Anglo-American adolescents provides a conceptual framework for those aspects of a psychosocial therapeutic model that enhances the appropriateness (and thus the effectiveness) of treatment for a Cuban immigrant adolescent population vis-à-vis an Anglo-American population. As part of a programmatic research effort to investigate the cultural characteristics of Cubans as well as developing and investigating the treatment of behavioral disorders in this population, a study of Cuban/Anglo-American adolescent value differences was conducted. The implications of these value differences for the appropriateness of psychosocial treatment models are discussed.

Theory

Clinical experience in the treatment program and a survey of the literature on cross-cultural comparisons of value orientations suggested that the theory of value orientations developed by Kluckhohn and Strodtbeck (1961) would provide a useful framework for contrasting cultural differences between Cuban immigrants and Anglo Americans. They postulated that to compare profiles between two cultures, it is necessary to delineate common human problems and to investigate the corresponding range of variations or ways of responding to these problems to the two dif-

ferent cultures. They describe five human problems common, in general, to all cultures. The solutions provided by each culture to these problems are indicative of world view or basic value orientations within that culture. From Kluckhohn and Strodtbeck (1961), the following definitions of the five basic areas of human problems and the range of possible solutions to these problems were derived:

1. *Human nature* orientation pertains to a society's perception of innate human qualities in terms of good and evil: (a) good—the human being is perceived as being basically good but corruptible; (b) evil—the human being is perceived as being basically evil but perfectible; (c) neutral—the human being is perceived as neither good nor evil and subject to influence.

2. *Person-nature* orientation refers to the perceived relationship of people to natural and environmental phenomena: (a) subjugation to nature—the person is helpless and at the mercy of nature's forces (worldly or other worldly); (b) mastery over nature—the person is seen as capable of controlling nature, mainly through technology; (c) harmony with nature—person and nature are one, working together in harmony.

3. *Activity* orientation refers to the nature of the behaviors through which a person is judged or judges himself or herself: (a) doing—the person is judged by what he or she achieves and emphasizes success-oriented activities usually including externally measurable activities; (b) being—this variation emphasizes activities that are an expression of existing desires (spontaneous expression), and activity is perceived existentially; (c) being in becoming—the emphasis in this variation is on meditation about one's self, which leads to understanding and self-development.

4. *Time* orientation refers to the meaning or emphasis placed on a particular time period: (a) past—the traditions of the past ought to be maintained or recaptured; (b) present—emphasis is on present time and problems; (c) future—emphasis is on a consideration of the future in solving present problems.

5. *Relational* orientation refers to the nature of a person's relation to other people:

(a) lineal—the way people relate to each other is determined by their relative positions within a hierarchy; (b) collateral—people's relations to each other are determined by a horizontal network. In this network all persons are at the same level and relate to each other as "equals" having a place in the network; and (c) individualistic—people relate to others autonomously, not by hierarchical or lateral networks.

Hypotheses

From anthropological and clinical impressions obtained through the treatment program, the following hypotheses were formulated:

1. Human nature orientation does not differ significantly for Cuban immigrants and Anglo Americans.

2. Person-nature orientation differs significantly between Cuban immigrants and Anglo Americans, with the former endorsing subjugation to nature and the latter, mastery over nature.

3. Activity orientation differs significantly between Cuban immigrants and Anglo Americans, with the former endorsing being and the latter endorsing doing as a preferred activity orientation.

4. Time orientation differs significantly between Cuban immigrants and Anglo Americans, with the former endorsing present and the latter endorsing future as a preferred time orientation.

5. Relational orientation differs significantly between Cuban immigrants and Anglo Americans, with the former endorsing lineality and the latter endorsing individualism as a preferred relationship style.

Method

Subjects

There were two samples in this study: Sample 1 was used in the development of the Value Orientations (VO) Scale; Sample 2 served primarily to test the value differences between Cuban immigrant and Anglo American adolescents. The 533 participants in the study were obtained from various educational institutions, such as high schools, junior colleges, universities, and continuing education centers; from social agencies, such as senior citizens activity centers;

and from other frequently used facilities, such as Cuban medical clinics. All facilities were located in the Greater Miami area.

Sample 1 consisted of 325 persons, including 120 (37%) males and 205 (63%) females. In terms of ethnic background, Sample 1 contained 220 (67.7%) Cuban immigrants, 65 (20.0%) Anglo Americans, 12 (3.7%) non-Cuban Latins, and 28 (8.6%) Black Americans. The average age of Sample 1 was 25.1 years, with a standard deviation of 12.1 and a range from 15 to 77 years.

Sample 2 was comprised of 208 persons, 81 (39%) males and 127 (61%) females, of whom 56 (27%) were Cuban immigrants and 152 (73%) were Anglo Americans. Since the majority of the clients in the treatment program referred to as "identified patients" and labeled as *in need of treatment* by their families were adolescents, Sample 2 was chosen to be representative of this sector of the population in treatment with respect to age. The average age of Sample 2 was 16.4 years, with a standard deviation of 1.4 and a range from 14 to 22 years.

Development of the Value Orientations Scale

Item construction. The first step in the development of the VO scale consisted of preparing an initial set of items reflecting the nature of the five human problems defined by Kluckhohn and Strodtbeck (1961) but in a context relevant to the target population. Each of the problem situations was followed by three statements presenting three possible alternative solutions. The final set of 22 problem situations¹ consisted of 9 relational, 4 human nature, 4 person-nature, 3 time, and 2 activity items. Two parallel forms were prepared. The first form was in Spanish. The second form was devised by translating the original set of items into English. The technique of back translation was used to insure the equivalence of the items (Brislin, 1970).

For each problem situation, the person was required to choose the solution considered best and the solution considered worst. The scores for the keyed responses were as follows: A response of *best* for an item was given a score of 3, a response of *worst* for an item was given a score of 1. If an item was not endorsed as either *best* or *worse*, it was assigned a score of 2. A response for each item could thus range from 1 to 3. Each alternative response was scored as a separate variable. Thus, three alternative responses in each of 22 problem situations produced 66 variables with a score of 1, 2, or 3 for each variable.

Scale construction. Rather than assuming that items should be combined as predicted by Kluckhohn and Strodtbeck's (1961) theory of value orientations, the items were submitted to an empirical test. Fol-

¹ Copies of the original Spanish and English versions of the Value Orientations Scale are available from the first author.

Table 1
Correlation Matrix of Four Factors

Factor	1	2	3
1			
2	.086		
3	.131	.160	
4	.047	-.102	-.012

lowing item construction, the scales were administered to Sample 1. The item responses of Sample 1 were factor analyzed, using an alpha solution and an oblique rotation (Harris-Kaiser, Type I).² Four interpretable factors³ emerged from the analyses, accounting for 14.52% of the total variance, with Factors 1, 2, 3, and 4 accounting for 5.31%, 3.50%, 3.21%, and 2.50% of the total variance, respectively. The factors, although obtained by oblique rotations, proved to be nearly independent of each other. Thus, for all practical purposes, the factor structure obtained can be said to be orthogonal. Table 1 presents the intercorrelations among the factors.

The following scale descriptions flow from the item loadings of each factor: Factor 1 is clearly a "relational factor," consistent with Kluckhohn and Strodtbeck's (1961) relational dimension. The items loading greater than .30 on this factor comprise the Relational scale. A high score discloses an individualistic value orientation in which the locus of responsibility for a person's behavior rests with the individual; "relationships are based on individual autonomy; reciprocal roles are based on recognition of the independence of interrelating members" (Papajohn & Spiegel, 1971, p. 260). A low score reflects a belief in lineality in which the locus of accountability is defined by the social structure; "relationships on a vertical dimension are hierarchically ordered; reciprocal roles are based on a dominance-submission mode of interrelationship" (Papajohn & Spiegel, 1971, p. 260).

Factor 2 is mixed, including primarily relational items in addition to person-nature, activity, and human nature items. The items loading above .30 on this factor comprise the Idealized Humanistic scale. A high score is an endorsement of idealized humanistic values, including a belief in collaterality, egalitarian social systems, and a growth-oriented lifestyle in search of harmony, peace, and spiritual development. A low score indicates low endorsement of these idealized humanistic values and greater personal concern.

Factor 3 is also a mixed variable on the perceived relationship between person and nature and their time orientation. The items loading higher than .30 on this factor comprise the Mastery/Future versus Subjugation/Present scale. A high score represents an engaging approach to life, with emphasis on planning for change at points in time extending away from present into future and a belief in the ability to

overcome the natural forces and harness them for human benefit. A low score reveals a fatalistic acceptance of life's circumstances and a belief that little can be done to counteract the forces of nature to which human beings are subjugated. Temporal focus is on the present, whereas the future is seen as being unpredictable.

Factor 4 is definitely related to the perception of human qualities and impulses. This is consistent with Kluckhohn and Strodtbeck's (1961) human nature dimension. Those items loading more than .30 on this factor comprise the Human Nature scale. A high score reflects a perception of human beings as basically selfish, malicious, and evil. A low score indicates a perception of human beings as basically good although corruptible.

Reliability and validity. Internal consistency or alpha coefficients were calculated for Sample 1 for each factor. Factors 1, 2, 3, and 4 yielded alpha coefficients of .89, .84, .76, and .72, respectively. These coefficients are within the acceptable standards for scales that have achieved internal consistency, thus insuring their satisfactory levels of reliability.

Factorial validity was obtained for the VO scale by ascertaining its internal statistical structure through factor analytic techniques. The factorial composition produced four orthogonal VO subscales with high internal consistencies, thus providing high factorial validity.

Results

Value Comparisons

To compare the value orientations of Cuban immigrants and Anglo Americans, the VO scale was administered to the persons in Sample 2. Their item responses were scored as described above in the item construction section. A monotone scaling model for unspecified distribution forms, also known as a linear model, was adopted to develop the VO scale (cf. Nunnally, 1967). In other words, scores were obtained for each of the four factorially derived subscales of the VO scale by algebraically summing the item scores of all of the items that loaded on each factor. Items

² A factor analysis was also conducted for Sample 1 excluding the Black American sample. The factor structure obtained for this analysis was nearly identical to the factor structure obtained for the full sample. Therefore, the factor analysis for the entire sample was used for this study.

³ The items loading greater than .30 on each of the factors, and their factor loadings are available from the first author. These items also comprise the Value Orientation scales.

Table 2
Value Comparisons: Means, Standard Deviations, and t Ratios for the Value Orientation Subscales

Orientation Subtests					
Factor	Cubans		Americans		<i>t</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Sample 1					
1	46.54	9.68	57.78	12.74	6.55***
2	52.67	10.92	50.30	10.25	-1.60
3	51.82	11.47	54.97	9.48	2.22*
4	47.06	9.94	45.44	9.95	-1.15
Sample 2					
1	47.5	10.1	50.9	9.7	2.13*
2	47.5	9.5	50.9	9.9	2.27*
3	46.7	9.9	51.3	9.9	2.92**
4	51.4	9.0	49.5	10.1	-1.32

Note. For Sample 1, $n = 211$ Cubans and 65 Americans; for Sample 2, $n = 56$ Cubans and 152 Americans. Standard scores are presented with a mean of 50 and a standard deviation of 10.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

loading positively on the factors were added, and items loading negatively were subtracted. Estimates of the internal consistency of the subscales were calculated for Sample 2. The alpha coefficients for Subscales (Factors) 1, 2, 3, and 4 were .76, .58, .51, and .46, respectively.

The Value Orientations Scale

The scores obtained by Cuban immigrants and Anglo Americans on the four factorially derived VO subscales were compared for Samples 1 and 2; t statistics were computed, and the significance of the obtained differences was determined using two-tailed tests of significance. Table 2 presents the means, standard deviations, and t ratios for the differences between the subscale scores. The subscale scores presented in Table 2 were transformed into standard scores, with a mean of 50 and a standard deviation of 10. Since Sample 1 was used to develop the VO scale, the results obtained with Sample 2 were used to test the hypotheses. An examination of Table 2 indicates that in Sample 2, the groups differ significantly for three of the four VO subscales.

As predicted in Hypothesis 1, there were no significant, $t(206) = -1.32$, differences

between Cuban immigrants and Anglo Americans along the Human Nature dimension (Factor 4).

The single largest difference, $t(206) = 2.92$, $p < .01$, was obtained for the Person-Nature and Time subscale (Factor 3). As predicted in Hypotheses 2 and 4, Anglo Americans tended to value *mastery over nature* and preferred to plan for the *future*, whereas Cubans tended to endorse a *subjugation to nature* orientation and a *present-time* orientation.

It was not possible to test Hypothesis 3 directly with respect to the differences in *activity* orientations between Cuban immigrants and Anglo Americans, since none of the factors included a sufficient number of activity items. Moreover, the small number of these items loading on the factors resulted from an artifact in the development of the original VO scale, which included only two human problems purporting to tap activity value orientations.

As expected from Hypothesis 5, Cuban immigrants and Anglo Americans differed significantly, $t(206) = 2.13$, $p < .05$, along the Relational subscale (Factor 1): Anglo Americans tended to value *individuality over lineality* in interpersonal relations, whereas the converse was true for Cubans.

A significant, $t(206) = 2.27$, $p < .05$, and unexpected difference emerged for the Idealized Humanistic Value subscale (Factor 2): Anglo Americans tended to endorse idealized humanistic values, whereas Cubans tended not to endorse these idealized humanistic values.

The scores obtained by Cuban immigrants and Anglo Americans on the four factorially derived VO subscales were also compared for Sample 1. An examination of Table 2 shows that in Sample 1, the groups differed significantly on two of the four VO subscales.

As with Sample 2, Cuban immigrants and Anglo Americans in Sample 1 (a) did not differ significantly, $t(274) = -1.15$, along the Human Nature dimension (Factor 4) as predicted in Hypothesis 1; (b) differed significantly, $t(274) = 2.22$, $p < .05$, in the Person-Nature and Time (Factor 3) subscale scores in the direction predicted by Hypotheses 2 and 4; and (c) differed significantly, $t(274) = 6.55$, $p < .001$, along the Relational subscale in the direction predicted by Hypothesis 5. Contrary to the findings obtained with Sample 2, Cuban immigrants and Anglo Americans in Sample 1 did not differ significantly, $t(274) = 1.60$, along the Idealized Humanistic Value subscale.

Activity Items

To test Hypothesis 3, the response scores obtained by Sample 2 Cuban immigrants and Anglo Americans on each item solution to the two activity problem situations were compared. The sample used in these comparisons included Sample 2 plus 120 additional high school students of Cuban or Anglo-American background. t statistics were computed, and the significance of the obtained differences were determined using a two-tailed test of significance. The results indicate that Cuban immigrants endorsed both items reflecting a "doing" orientation, $t(326) = 5.55$, $p < .0005$, $t(326) = 2.41$, $p < .02$, significantly more frequently than Anglo Americans, whereas Anglo Americans endorsed both items reflecting a "being" orientation significantly more frequently than Cuban immigrants, $t(326) = 3.11$, $p < .002$; $t(326) = 1.98$, $p < .05$. There were no differences between Cuban immi-

grants and Anglo Americans in their endorsement of the two items indicative of a being-in-becoming orientation, $t(326) = .80$; $t(326) = .21$. These findings are contrary to the prediction in Hypothesis 3.

The item solutions to the two activity problem situations provided by Sample 1 Cuban immigrants and Anglo Americans were also compared using t statistics and two-tailed tests of significance. Again as with Sample 2, Sample 1 Cuban immigrants endorsed both items reflecting a doing orientation significantly more frequently than Anglo Americans, $t(274) = 2.02$, $p < .05$; $t(274) = 2.02$, $p < .05$; Anglo Americans tended to endorse both items reflecting a being orientation more frequently than Cuban immigrants, $t(274) = 1.85$, $p < .07$; $t(274) = 1.79$, $p < .08$; and, there were no significant differences between the groups in their endorsement of "being-in-becoming" items, $t(274) = .17$, $t(274) = .01$.

Factor Structure of Sample 2

To ascertain the generalizability of the VO scale to Sample 2, the item responses of Sample 2 were also factor analyzed, using an alpha solution and an oblique rotation (Harris-Kaiser, Type I). Three of the four factorially derived VO subscales obtained from Sample 1 were discernible in the factors that emerged from Sample 2.

Factor 1, comprising the Relational subscale, proved to be the strongest factor in both Samples 1 and 2, accounting in each case for the largest proportion of the total factor variance. Of the 17 items loading on the Relational subscale, 14 items (82%) also loaded on the first factor of Sample 2.

Factor 2, comprising the Idealized Humanistic Value subscale, emerged as the second strongest factor in both Samples 1 and 2. Of the 13 items loading on the Idealized Humanistic Values subscale, 7 items (54%) also loaded on the same factor for Sample 2. An apparent difference between these factors was observed, however. For Sample 1, the factor that emerged was essentially unipolar, measuring low to high idealized humanistic values. The factor that emerged from Sample 2 was clearly bipolar, ranging from idealized

humanistic values on the one hand to a "life is a jungle" value orientation on the other, with an emphasis on the evil qualities of people and the need for self-protection as a survival measure.

Factor 3 of Sample 1, comprising the Mastery over Nature/Future Time versus Subjugation to Nature/Present Time subscale, emerged also as a factor in Sample 2. Six of the nine items (67%) of this subscale loaded on a factor of Sample 2. In this instance, the factor that emerged for Sample 2 was clearly a mastery over nature/future time versus subjugation to nature/present time factor.

It was not possible to identify a factor for Sample 2, which appeared comparable to the Human Nature subscale measured by Factor 4 of Sample 1.

Discussion

The VO scale was developed using Sample 1, and the hypotheses were tested using Sample 2. However, to ascertain the stability of the results across both samples, each sample was factor analyzed separately, and the value comparisons between Cuban immigrants and Anglo Americans were also conducted separately for each sample. The first two factors emerged strongly in the factor structure of both samples; the third factor emerged from Sample 1 and was replicated partially in Sample 2; and the fourth and weakest factor that emerged from Sample 1 was not identifiable in the factor structure of Sample 2. These differences in factor structure are not surprising, since the factors with the highest eigenvalues replicated better across samples, and the factor with the lowest eigenvalue failed to replicate across samples.

With one exception, the differences in value orientations between Cuban immigrants and Anglo Americans held for both samples. The only exception occurred along the idealized humanistic value dimension. Whereas Sample 2 Anglo Americans were significantly higher than Sample 2 Cuban immigrants on this value, the same two groups in Sample 1 did not differ significantly along this dimension. Since Sample 2 subjects were younger (M age = 16.4) than Sample 1 subjects (M age

= 25.1), it is suggested that the relative shift on idealized humanistic values between the groups in the two samples may have resulted from the differences in age between the samples.

It is interesting to note that Cuban immigrants and Anglo Americans may diverge on many culturally related variables other than nationality. For example, these groups may vary on religious affiliation, child-rearing customs, family structure, and psychological variables such as need for approval, locus of control, and field dependence. The present study did not attempt to control for these variables. To have singled out anyone or a combination of these variables for analyses would have been artificial, since, in fact, these variables and many others contribute to the differences in basic value orientations observed between the two cultural groups examined.

The differences between the two cultural groups may have been caused, however, by variables that are not necessarily culture related. For example, Casavantes (1970) argued that Mexican Americans value the present only as a function of their lower socioeconomic status and not as a cultural value. Since socioeconomic status data were available for the subjects in Sample 2, the Cuban immigrants and Anglo Americans of Sample 2 were compared along this variable, and they were found not to differ significantly. Hence, the findings of the present study do not appear to have been caused by socioeconomic differences as Casavantes would suggest.

Future research should address cultural differences among Cuban subgroups. Differences may exist, for example, between the sexes, among age groups, or between clinical and nonclinical samples. These differences between Cuban subgroups may also have important clinical implications.

Clinical Implications

The differences in basic value orientations between Cuban immigrant and Anglo-American adolescents may have implications for the delivery of mental health services to these populations. If these value orientations are indeed as basic as Kluckhohn and Strodtbeck (1961) postulated, then they must also have

implications for personality and psychosocial development, a notion derived from the work of Papajohn and Spiegel (1971) and Ramirez and Castañeda (1974) among others.

As suggested earlier, clients with specific psychosocial characteristics require treatment approaches matched to their idiosyncratic styles. Following this premise, it would seem that to achieve desired psychotherapeutic outcomes with clinical Cuban populations, it is necessary to identify treatment models that are specifically matched to the culturally determined characteristics arising from the value structure of this population. Therefore, the Cuban adolescents' preference for lineality, subjugation to nature, present time, and doing orientations as well as their low endorsement of idealized humanistic values must be taken into consideration when designing a psychosocial service delivery system for them. Many traditional Anglo-American treatment services are based on a model of a growth-oriented, self-actualizing individual who is ready to take control over his or her own destiny. In contrast, clinical experience at the Spanish Family Guidance Clinic suggests that the provider of psychosocial treatment services to the Cuban immigrant must be ready to take charge of the therapist-client relationship, to validate hierarchical structures in the client's life context, and to intervene on behalf of the client within the client's life context to restore ecological order.

The most important feature of a psychosocial treatment model that is sensitive to the cultural characteristics of the Cuban immigrants is to validate their preference for a lineal style of relationships. This relationship style may receive the support of the therapist in various phases of the treatment. First, the therapist must relate to the client hierarchically, recognizing that the therapist's role is perceived by the client as a position of authority. With this recognition, the therapist assumes responsibility and further takes charge of the therapist-client relationship. Second, the therapist validates the young Cuban client's preference for lineality by enlisting in the treatment process the naturally occurring hierarchical systems in the client's life context. Clearly, the most significant na-

turally occurring hierarchical system is the family. Other authority figures such as teachers, school counselors, and even probation officers may also be important and may need to be included in the therapeutic plan. Many instances of dysfunction in young clients are also accompanied by the breakdown of the lineal structural relational patterns within the family. This is frequently manifested in the young person's open rejection of the parents' executive role in the family. Interestingly enough, even in these instances, clinical experience suggests that desired therapeutic outcomes are reached most expediently by restoring the lineal-hierarchical relational structure in the family. Once the family's natural lineal milieu is restored, and the parents' role as the family's executive system is reaffirmed, then the family is taught the skills necessary to negotiate the youngster's differentiation (transition from lineality to individuality) from the family. The lineal family structure is supported, so that within this basic culturally sanctioned framework, the process of negotiating the youngster's differentiation from the family may take place.

In preparing a psychosocial treatment plan for a Cuban immigrant client, the therapist should also consider the Cuban's sensitivity to environmental social pressures. The high levels of need for approval of Cubans (Tholen, 1974) and field dependence of Latins (Ramirez & Castañeda, 1974) in general have been documented. Because of the strong influence of environmental social pressures on the Cuban client's well-being, it becomes particularly important that the etiology of psychosocial dysfunctions be conceptualized within an ecological framework (Auerswald, 1971), since ecological theory takes into consideration the effects of the interaction between client and psychosocial systems on the client's functioning. Further, as the present study indicates, Cuban clients tend to perceive themselves as unable to control or modify their environmental circumstances (see also Santisteban, 1975). For this reason, when environmental pressures or tensions seem to be a source of client dysfunction, as is frequently the case (Scopetta, King, & Szapocznik, Note 1), it is necessary that therapeutic interventions re-

structure the interactions of the client with his/her environment when these are sources of client functional impairment (Aponte, 1974).

The treatment of the Cuban client must also be present oriented. The Cuban client is usually mobilized for treatment by the onset of a crisis (Scopetta et al., Note 1) and expects the therapist to provide immediate problem-oriented solutions to the crisis situations. In general, the therapist must develop a treatment model that capitalizes on crises to promote personal growth and the reorganization of interpersonal relations. Further, to use maximally this characteristic of the Cuban population, the culturally sensitive therapist is not only cognizant of how to use crises to promote growth but also knows how to create them for the same purpose.

It will be recalled that young Anglo Americans are more likely than Cubans to endorse idealized humanistic values (Factor 2). Since these findings were unexpected, their clinical implications are as yet not clear. Nevertheless, these findings suggest that young Cuban immigrants are less likely than their Anglo-American counterparts to value relationships based on goals of the laterally extended group, and thus to be mobilized in treatment by peer pressure groups. The findings also suggest that young Cuban immigrants are less likely than their Anglo American counterparts to be motivated in treatment by a search for personal and spiritual growth. In fact, clinical experience suggests that the Cubans are motivated in treatment by concrete and obtainable objectives. Consistent with this interpretation, young Cuban immigrants were found to endorse a doing activity orientation, whereas their Anglo American counterparts preferred a being activity orientation.

With recognition of the culturally determined characteristics of the Cuban population, the Spanish Family Guidance Clinic in the Department of Psychiatry of the University of Miami School of Medicine explored a variety of treatment approaches. Among these, one treatment model seemed particularly appropriate, ecological structural family therapy, first proposed by Aponte (1974), who based his treatment model on Auers-

wald's (1971) concepts of ecological therapy and Minuchin's (1974) structural family therapy. The approach of these therapists seems to be particularly appropriate for the treatment of Puerto Ricans (e.g., Minuchin, Montalvo, Guernsey, Rosman, & Schumer, 1967).

Ecological structural family therapy as adopted at the Spanish Family Guidance Clinic is based on therapeutic assumptions that are matched to the value characteristics of the population of Cuban adolescents. Within this approach, the therapist relates hierarchically to the client and works to restore the hierarchical structure in the client's family. The therapist considers the ecological factors impinging on the client and actively intervenes to remediate detrimental ecological relationships based on the notion that the client lacks the orientation to do so unassisted. The therapist is present oriented and intervenes to manipulate existing dysfunctional interactional patterns (cf. Minuchin, 1974) within the family and between the family and its environment. And, finally, consistent with a doing activity orientation, the client is motivated for treatment through the use of concrete and obtainable objectives.

Further studies are under way to test the effectiveness of ecological structural family therapy in the treatment of psychosocial dysfunctions, including drug and alcohol abuse, with a population of Cuban immigrants.

The procedure outlined in this article may have broad implications for the development of culturally specific psychosocial treatment models. It is suggested that this procedure may be applicable to the development of culture specific treatment for other cultural groups. In fact, the VO scale may be used to ascertain basic cultural characteristics in client populations or for specific clients in treatment. Based on the findings obtained with the VO scale, it is then possible to identify treatment features that "match" the individual client or clients population's basic value orientations.

Reference Note

1. Scopetta, M. A., King, O. E., & Szapocznik, J. *Relationship of acculturation, incidence of drug abuse, and effective treatment for Cuban Americans.*

(Contract 271-75-4136) Research proposal, National Institute on Drug Abuse, June 1975.

References

- Al-Issa, I. Culture and symptoms. In C. G. Costello (Ed.), *Symptoms of psychopathology: A handbook*. New York: Wiley, 1970.
- Aponte, H. J. Psychotherapy for the poor: An ecological approach to treatment. *Delaware Medical Journal*, March 1974, 1-7.
- Auerswald, E. H. Families, change, and the ecological perspective. *Family Process*, 1971, 10, 263-280.
- Bergin, A. E. The evaluation of therapeutic outcomes. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change*. New York: Wiley, 1971.
- Berry, J. W., & Annis, R. Acculturative stress: The role of ecology, culture, and differentiation. *Journal of Cross-Cultural Psychology*, 1974, 5, 382-406.
- Brislin, R. W. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1970, 1, 185-216.
- Casavantes, E. Pride and prejudice: A Mexican-American dilemma. *Civil Rights Digest*, 1970, 3, 22-27.
- Kiesler, D. J. A grid model for theory and research in the psychotherapies. In L. D. Eron & R. Callahan (Eds.), *The relationship of theory to practice in psychotherapy*. Chicago: Aldine, 1969.
- Kiesler, D. J. Experimental designs in psychotherapy research. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change*. New York: Wiley, 1971.
- Kluckhohn, F. R., & Strodtbeck, F. L. *Variations in value orientations*. Evanston, Ill.: Row, Peterson, 1961.
- Ladner, R., Page, W. F., & Lee, M. L. Ethnic and sex effects on emergency ward utilization for drug-related problems: A linear model analysis. *Journal of Health and Social Behavior*, 1975, 16, 315-325.
- Mezey, A. G. Psychiatric aspects of human migrations. *International Journal of Social Psychiatry*, 1960, 5(4), 245-260.
- Minuchin, S. *Families and family therapy*. Cambridge, Mass.: Harvard University Press, 1974.
- Minuchin, S., Montalvo, B., Guernsey, B., Jr., Rosman, B. L., & Schumer, F. *Families of the slums*. New York: Basic Books, 1967.
- Nunnally, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.
- Padilla, A. M., & Ruiz, R. A. *Latino mental health: A review of the literature* (Publication No. 1724-00317). Washington, D.C.: U.S. Government Printing Office, 1973.
- Papajohn, J. C., & Spiegel, J. P. The relationship of culture value orientation change and Rorschach indices to psychological development. *Journal of Cross-Cultural Psychology*, 1971, 2, 257-272.
- Paul, G. L. Behavior modification research: Design and tactics. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- Ramirez, M., & Castañeda, A. *Cultural democracy, bicultural development and education*. New York: Academic Press, 1974.
- Santisteban, D. *Locus of control as a cross-cultural variable*. Unpublished master's thesis, University of South Florida, 1975.
- Tholen, J. F. *An interactive approach to the study of outcome in group counseling: Matching conceptual level with degree of structure*. Unpublished master's thesis, University of Miami, 1974.
- Weidman, H. H. Concepts as strategies for change. *Psychiatric Annals*, 1975, 5, 312-314.

Received July 18, 1977 ■

Personality Characteristics of Long-Term Recovered Alcoholics: A Comparative Analysis

William M. Kurtines, Leah R. Ball, and Gloria H. Wood
Florida International University

This study reports data on the personality characteristics of alcoholics in two stages of recovery (short and long term). Three samples (total $N = 183$) were used in this study: (a) 60 newly recovered alcoholics (30 males and 30 females with at least 3 weeks but less than 4 months of sobriety); (b) 62 long-term recovered alcoholics (31 males and 31 females with a minimum of 4 years of continuous sobriety and a mean length of sobriety of 8.9 years); and (c) 61 nonalcoholic controls (30 males and 31 females who reported moderate to infrequent or no use of alcohol). All subjects were administered the California Psychological Inventory and a biographical data sheet. A multivariate analysis of variance was used to test the significance of group differences on the personality variables, and a multiple discriminant analysis was conducted to determine the most discriminating dimensions for differentiating among the three groups. The results of the analysis clearly indicate the existence of differential patterns of psychological adjustment at each stage of recovery.

Early research on the personality characteristics of alcoholics focused on the identification of a single alcoholic personality type. The results of this research effort were generally disappointing (Sutherland, Schroeder, & Tordella, 1950; Syme, 1957), and the need for alternative conceptualizations are obvious. Recent research studies have tended to focus on the identification and classification of personality patterns common to alcoholics (Goldstein & Linden, 1969; Lawlis & Rubin, 1971; Nerviano, 1976; Nerviano & Gross,

1973; Partington & Johnson, 1969; Skinner, Jackson, & Hoffman, 1974; Whitelock, Patrick, & Overall, 1971), and the results have led to a more realistic reappraisal of alcoholics as a heterogeneous treatment population (Nerviano, 1976). In general, these studies have used multivariate data analysis procedures to identify distinctive personality profiles among newly recovered alcoholics or alcoholic patients undergoing treatment. Skinner et al. (1974), for example, used a multivariate classification strategy and the Differential Personality Inventory (DPI; Jackson & Messick, 1970) to identify eight modal personality profiles among male alcoholic psychiatric patients. A more recent study by Nerviano (1976), using the Personality Research Form (PRF; Jackson, 1967), used a multivariate approach to identify seven common personality patterns among male alcoholics undergoing treatment in a Veterans Administration hospital.

Overall, considering the variety of assessment devices and methodological approaches used in these studies, there has been a remarkable convergence of results. Skinner, Reed, and Jackson (1976), for instance, have

The authors would like to express appreciation to Wayne F. Cascio, Harrison G. Gough, Dennis Murphy, and Vincent J. Nerviano for their helpful comments on an earlier draft of this paper. We would also like to thank the following individuals for their kind assistance in the gathering of the data: David Altschuler, Director Residential Services, Comprehensive Alcohol Program, Dade County, Florida; Ross A. Garnsey, Director and Founder, Bethesda Manor, Catholic Service Bureau, Dade County, Florida; and Lester Bashara for the use of the facilities of the Miami Church of Religious Science.

Requests for reprints should be sent to William M. Kurtines, Department of Psychology, Florida International University, Tamiami Trail, Miami, Florida 33199.

demonstrated the feasibility of generalizing the modal personality profiles derived from alcoholics (cf. Skinner et al., 1974) to other psychiatric samples. Moreover, Skinner et al. (1976) observed that two particular profiles (Type I: defensive-repressive, and Type II: impulsive-socially deviant) were notably pervasive across all of the samples used in their study, a finding that is strikingly consistent with those reported by Nerviano (1976). The two most frequently occurring personality patterns in the sample of alcoholics used by Nerviano (1976) were Type A (high impulse control, low autonomy) and Type B (extremely low impulse control).

In view of the substantial body of research currently available on the personality patterns of alcoholics in the initial stages of recovery, it is surprising to find that to date, little research has been conducted with long-term recovered alcoholics. There are, for example, no readily available studies on the personality characteristics of these individuals. The paucity of research on long-term recovered alcoholics is particularly striking in light of the generally acknowledged difficulty in working with alcoholics as a treatment population (Huber & Danahy, 1975; Tamerin & Neumann, 1974). The lack of research in this area may be due, in part, to the difficulty in obtaining data on such persons. Long-term recovered alcoholics are not typically identified with treatment programs and consequently are not readily available to the interested researcher. Moreover, because of the social, personal, and occupational stigma associated with alcoholism, the recovered alcoholic usually seeks to preserve anonymity. Thus, although the research literature contains valuable information about the personality characteristics of newly recovered alcoholics, there is a conspicuous lack of data on the characteristics of individuals who have been able to maintain long-term sobriety. Such data would provide useful clinical information concerning the specific types of personality characteristics associated with long-term recovery among alcoholics as a treatment population.

The purpose of this study was to identify differential personality characteristics associated with both short- and long-term recovery

among alcoholics. The samples used in the study consisted of alcoholics in two stages of recovery and a sample of non-alcoholic controls. The California Psychological Inventory (CPI; Gough, 1957) and a biographical data sheet were administered to all subjects, and multivariate data analysis techniques were used to determine differential personality patterns for the three samples. Since the research was intended to identify factors related to positive adjustment as well as clinical symptomatology, the CPI appeared to be a logical choice as an assessment device. The CPI is designed for use with a general population and yields information on the adequacy of interpersonal as well as intrapersonal functioning.

Method

Subjects

Three samples (total $N = 183$) were used in this study. Group 1, the short-term recovered alcoholics, consisted of 60 "dry" alcoholics with at least 3 weeks but less than 4 months of sobriety (30 males and 30 females). The majority were residents in halfway houses in the Miami, Florida, area, and 19 were new members from Alcoholics Anonymous (AA). Group 2, the long-term recovered alcoholics consisted of 62 sober alcoholics (31 males and 31 females) with a *minimum* of 4 years of continuous sobriety, the mean length of sobriety being 8.9 years. All participants in Group 2 were active members of AA. Group 3, the nonalcoholic controls, consisted of 61 adult subjects (30 males and 31 females) chosen at random from the Miami, Florida, area. All members of the control group reported moderate to infrequent or no use of alcohol.

All subjects were between the ages of 30 and 65. From the information available, the sober alcoholic and the nonalcoholic control groups appeared to be matched in terms of socioeconomic status. The dry alcoholics, on the other hand, ranked somewhat lower in terms of the traditional indices of socioeconomic status (e.g., occupation, income, etc.). These differences, however, appear to be related to the duration of their alcoholism and the recency of their recovery more than their actual status. The mean ages of the groups were as follows: dry alcoholics, 46 years ($SD = 8.1$); sober alcoholics, 49 years ($SD = 11.1$); nonalcoholic controls, 49 years ($SD = 8.4$).

Procedure

The CPI and a biographical data sheet were administered to all participants individually or in small groups. For the sober alcoholics, most of the ad-

ministration was conducted at informal meetings of active members of AA in the Miami area. The CPI was administered to the dry alcoholics, with the exception of the 19 new members of AA, on site at the various halfway houses in the Miami area. For the nonalcoholic controls, the CPI was administered in small groups to adult volunteers drawn from a wide variety of occupations (firefighter, meat cutter, pilot, postal worker, teacher, homemaker, etc.) and social backgrounds. The CPI was scored for the 18 standard scales plus an additional scale developed by Hogan (1969).

Results

A 2×3 multivariate analysis of variance (Clyde, 1969), using raw scores on the CPI scales as dependent variables, was conducted to determine the existence of sex and group differences on the CPI profiles. A multivariate test of significance, using Wilk's lambda

criterion, indicated a significant main effect due to sex, $F(19, 159) = 9.64, p < .001$, but no significant Sex \times Group interaction, $F(38, 318) = .899, p > .05$. An examination of the univariate F tests for the main effect of sex further indicated that males scored significantly higher on Well-Being (Wb), $F(1, 177) = 4.64, p < .05$; and females scored higher on Achievement via Independence (Ai), $F(1, 177) = 5.30, p < .01$, and Femininity (Fe), $F(1, 177) = 149.36, p < .001$.

The multivariate analysis of variance also yielded a significant main effect for group, $F(38, 318) = 3.37, p < .001$. More importantly, the pattern and direction of the significant differences obtained for the univariate F tests conducted for the individual scales provided strong evidence for the existence of

Table 1
Means, Standard Deviations, F Ratios, and Scheffé's S Method for the Samples Listed, Using the California Psychological Inventory and the Social Maturity Index (SMI)

Scale	Group 1 ^a		Group 2 ^b		Group 3 ^c		F	Scheffé		
	M	SD	M	SD	M	SD		1 vs. 2	1 vs. 3	2 vs. 3
Do	24.2	6.9	27.0	7.1	29.9	6.6	10.2*	—	.001	—
Cs	17.4	4.9	19.1	3.9	21.4	3.9	12.8*	—	.001	.01
Sy	22.0	6.1	22.4	5.6	26.0	5.4	9.0*	—	.001	.01
Sp	33.5	7.5	34.8	6.0	36.2	6.8	2.3	—	—	—
Sa	19.8	4.7	21.4	4.1	22.1	3.4	5.0*	—	.05	—
Wb	32.1	6.8	35.3	4.9	36.5	5.5	9.3*	.05	.001	—
Re	24.3	5.3	28.2	5.6	31.0	4.9	23.1*	.001	.001	.05
So	28.0	5.0	30.2	5.2	35.0	4.6	27.2*	.05	.001	.001
Sc	24.6	8.5	27.1	7.0	30.5	7.2	9.5*	—	.001	.05
To	18.1	5.9	21.6	5.0	23.0	5.0	13.6*	.01	.001	—
Gi	13.0	5.5	15.7	6.0	19.8	6.9	18.4*	—	.001	.01
Cm	25.0	3.2	25.7	2.0	25.5	2.4	1.5	—	—	—
Ac	22.8	5.4	25.3	5.1	28.8	4.4	22.1*	.05	.001	.01
Ai	18.0	4.9	20.8	4.1	21.0	4.4	9.0*	.01	.01	—
Ie	34.1	7.6	36.7	6.1	39.4	5.9	9.9*	—	.001	.05
Py	10.1	3.4	11.6	2.9	12.5	2.9	10.1*	.05	.001	—
Fx	9.8	4.8	10.0	4.3	9.9	4.3	.0	—	—	—
Fe	20.1	4.9	20.9	3.8	19.7	4.6	1.1	—	—	—
Em	19.3	5.2	20.9	4.9	23.5	4.9	11.4*	—	.001	.05
SMI	45.4	3.7	47.2	3.2	49.4	3.3	22.1*	.05	.001	.05

Note. Do = Dominance; Cs = Capacity for Status; Sy = Sociability; Sp = Social Presence; Sa = Self-Acceptance; Wb = Well-Being; Re = Responsibility; So = Socialization; Sc = Self-Control; To = Tolerance; Gi = Good Impression; Cm = Communality; Ac = Achievement via Conformance; Ai = Achievement via Independence; Ie = Intellectual Efficiency; Py = Psychological-mindedness; Fx = Flexibility; Fe = Femininity; Em = Empathy.

^a Dry alcoholics ($n = 60$).

^b Sober alcoholics ($n = 62$).

^c Nonalcoholic controls ($n = 61$).

* $p < .001$.

distinctive personality profiles for each of the groups. Fifteen of the 19 univariate F tests were significant at the .001 level, and 31 of the 57 post hoc comparisons were significant at more than the .05 level. Table 1 presents the means, standard deviations, and F ratios for all of the CPI scales for the three samples. Table 1 also presents the post hoc comparisons between means using Scheffé's S method (Kirk, 1968).

The single most notable finding reported in Table 1 is the large number of significant differences between the personality profiles of all three groups. As can be seen from Table 1, the personality profiles of alcoholics in two stages of recovery (i.e., newly recovered and long-term recovered) differed significantly from nonalcoholic controls and from each other. The results of the group comparisons can be summarized as follows: First, the profile of the sober alcoholics differed significantly from that of the nonalcoholic controls on the following nine scales: Capacity for Status (Cs; $p < .01$), Sociability (Sy; $p < .01$), Responsibility (Re; $p < .05$), Socialization (So; $p < .001$), Self-control ($p < .05$), Good Impression ($p < .01$), Achievement via Conformance (Ac; $p < .01$), Intellectual Efficiency ($p < .05$), and Empathy (Em; $p < .05$). Moreover, the nonalcoholic controls exhibited a more normal profile, scoring higher than the sober alcoholics on all nine scales. Second, the dry alcoholics had an *extremely* depressed profile in comparison to that of the nonalcoholic controls. The dry alcoholics scored significantly lower on all but the following four scales: Social Presence, Communality, Flexibility, and Fe. Furthermore, 13 of the 15 significant differences were in excess of .001. Finally, in comparison to the dry alcoholics, the sober alcoholics in general exhibited a more elevated profile, scoring higher than the dry alcoholics on the following seven scales: Wb ($p < .05$), Re ($p < .001$), So ($p < .05$), Tolerance ($p < .05$), Ac ($p < .05$), Ai ($p < .01$), and Psychological-mindedness ($p < .05$).

The second phase of the analyses was designed to determine the personality dimensions most useful in differentiating among the groups. A multiple discriminant analysis was conducted using group membership as a

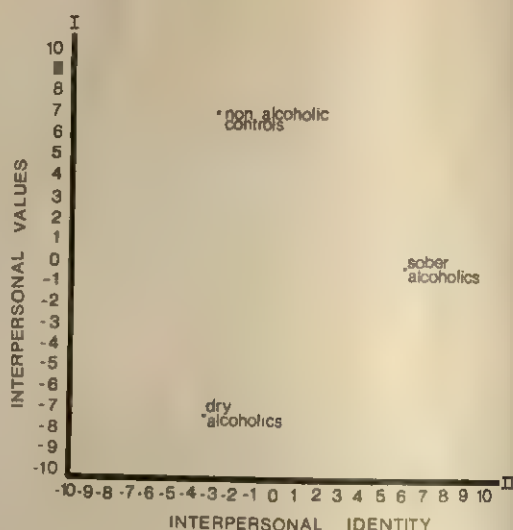


Figure 1. Group centroids for discriminant functions I and II.

criterion variable and CPI scale scores as predictor variables (Cooley & Lohnes, 1971). The analysis yielded two significant discriminant functions; and a measure of overall group differentiation, Wilk's lambda, indicated that both functions significantly discriminated between groups ($p < .001$). The centroids of the three groups are plotted in Figure 1, with the first significant discriminant function serving as the ordinate and the second discriminant function serving as the abscissa. These group centroids are the mean scores for the individuals within each group for the two significant functions.

The first and most important dimension for differentiating between groups, which accounted for 69.6% of the between-group variance, was a bipolar dimension defined on the positive end by So (.56), Re (.41), and Em (.37) and defined on the negative end by Sy (-.37). This dimension, which was labeled *interpersonal values*, appears to reflect a socially mature, empathic, and articulated value system versus a primarily affiliative, other-directed value orientation. From the post hoc comparisons presented in Table 1, it can be seen that the dry alcoholics had the lowest scores on So, Re, and Em (the CPI "moral" scales) and Sy (a measure of interpersonal affiliation). More importantly, as can be seen from Figure 1, the first func-

tion differentiates among all three groups, with the nonalcoholic controls scoring highest and the dry alcoholics scoring lowest.

The second significant function, which accounted for 30.4% of the between-group variance, is a dimension of general profile elevation. Nine of the CPI scales loaded above .35 on this function, whereas only four loaded above .35 on the first function. This dimension of general profile elevation is defined primarily by Self-Acceptance (Sa) (.58) and Wb (.56) on the positive end and by Sy (-1.36) on the negative end. High scores on Sa indicate the presence of a positive self-image and high self-esteem; high scores on Wb indicate a sense of well-being and a freedom from excessive health concerns. Low Sy scores, on the other hand, indicate social detachment, passivity, and generally poor interpersonal functioning. This dimension, which was labeled *interpersonal identity*, appears to reflect a relatively conflict-free and comfortable sense of personal worth versus uncertainty in self-evaluation resulting in a need for interpersonal supports. The significance of this dimension can be seen from the pattern of centroid plots displayed in Figure 1. The sober alcoholics scored highest on this function, whereas the dry alcoholics and the nonalcoholic controls obtained about the same mean centroid, suggesting that long-term recovered alcoholics display a unique psychological adjustment profile. On this dimension, long-term recovered alcoholics appeared to be relatively nonneurotic but moderately socially maladjusted. Confirmation of this interpretation can be found in Table 1. The sober alcoholics did not differ significantly from the nonalcoholic controls on Sa or Wb, but they scored significantly lower than the controls on Cs and Sy.

For the final phase of the analyses, subjects in all three groups were scored for a CPI-based "social maturity index" developed by Gough (1966). The social maturity index was originally defined by comparing the responses of a large sample of delinquents and nondelinquents on the CPI and developing a six-variable regression equation to distinguish between the groups. Subsequent research (Gough, De Vos, & Mizushima, 1968; Hogan, Mankin, Conway, & Fox, 1970) has estab-

lished the utility of the social maturity index as a measure of antisocial tendencies. The constant and the weights for the equation have been adjusted so that the mean score on the index in a normal population will be 50.0. In the original sample, the nondelinquent mean was 50.4 and the mean delinquent score was 42.7. Mean scores for the dry alcoholics, sober alcoholics, and nonalcoholic controls used in this study were, respectively, 45.4 ($SD = 3.7$), 47.2 ($SD = 3.2$), and 49.4 ($SD = 3.3$). A one-way analysis of variance, using social maturity scores as the dependent variable, was highly significant, $F(2, 180) = 22.05$, $p < .001$. Post hoc comparisons, using Scheffé's S method, were significant at or above the .05 level for all paired comparisons.

Discussion

This study presents data on the personality characteristics of alcoholics in two phases of recovery. A multivariate analysis of variance was used to determine group differences on the personality variables, and a multiple discriminant analysis was utilized to identify the most useful dimensions for differentiating between both groups of alcoholics and a group of nonalcoholic controls. The overall results of the analyses indicate the existence of differential patterns of adjustment and coping strategies at each stage of recovery.

The personality profile of the alcoholics in the initial stages of recovery showed a marked similarity to common personality patterns identified in previous research. The profile of the newly recovered alcoholics can be summarized as follows: First, the dry alcoholics exhibited an extremely depressed profile, indicating a generally poor level of adjustment. Second, this group was characterized by strong antisocial tendencies and impulsiveness, scoring significantly lower on the social maturity index than both other groups. Third, the dry alcoholics appeared to exhibit a sense of interpersonal inadequacy, low self-esteem, and feelings of guilt and self-blame. This overall pattern is notably consistent with several of the modal personality profiles reported in previous research with alcoholic patients. In particular, the CPI profile of the newly recovered alcoholics in this study bears

a strong resemblance to the Type II personality (impulsive-socially deviant) identified by Skinner et al. (1976) using the DPI, and to the Type B personality (extremely low impulse control) identified by Nerviano (1976) using the PRF and the Sixteen Personality Factor Questionnaire. The convergence of results obtained with such a wide range of assessment devices provides strong evidence for the pervasiveness of this personality type among alcoholics at this stage of recovery.

The long-term recovered alcoholics, on the other hand, displayed a unique personality profile, differing from both the newly recovered alcoholics and the nonalcoholic controls. The profile of the sober alcoholics can be described as follows: First, the sober alcoholics displayed a generally more elevated profile than the dry alcoholics but significantly less elevated than the controls, a finding which suggests that the maintenance of long-term sobriety is associated with an overall improvement in general level of adjustment. Second, the sober alcoholics appeared to be moderately undersocialized in terms of intrapersonal values. As a group, they obtained a midrange score on the social maturity index. Moreover, they scored significantly higher on Re and So than the dry alcoholics but lower than the nonalcoholic controls, suggesting that recovery is significantly but not exclusively related to a restructuring of intrapersonal values. Third, the sober alcoholics displayed a pattern of poor interpersonal functioning similar to that of the dry alcoholics. As a group, they scored significantly lower on Cs and Sy than the controls, indicating a sense of interpersonal inadequacy and social inhibition. This generally poor level of interpersonal functioning stands in sharp contrast to their generally good level of intrapersonal functioning. They did not differ significantly from the controls on CPI scales concerned with intrapersonal functioning (Sa and Wb), but the dry alcoholics did. This finding suggests that long-term recovered alcoholics, although socially inhibited, are relatively self-accepting, have a strong sense of well-being, and are free from excessive health concerns.

The findings reported in this study, although tentative, appear to have implications for both treatment intervention and the identification of areas of much needed research. Current research efforts have tended to focus on the identification of distinctive subgroups of personality types among diagnosed alcoholics and have emphasized the heterogeneous nature of the treatment population. The notable success of such research efforts has led to valuable suggestions concerning treatment intervention. Nerviano (1976), drawing on his extensive research with alcoholic patients, has suggested that treatment type and modality be adjusted to the individual alcoholic on the basis of identified personality typologies. The results reported here, while not directly concerned with the identification of personality subtypes, are consistent with this previous research and would further extend this suggestion to include a consideration of stage of recovery as an important factor in the determination of treatment modality and type. For example, one of the clinically most significant findings of this study was that the overall personality profile of the long-term recovered alcoholics differed significantly from both the newly recovered alcoholics and the nonalcoholic controls, suggesting a distinctive pattern of psychological adjustment among long-term recovered alcoholics. The most discriminatory dimension for differentiating between long-term recovered alcoholics and the other two groups was defined by (a) the general level of profile elevation, (b) an absence of psychoneurotic tendencies, and (c) poor interpersonal functioning. This finding would seem to strongly suggest that treatment intervention be adjusted to include stage of recovery as well as overall personality profile. Finally, the results reported here point to the need for more extensive and systematic research on the personality characteristics of long-term recovered alcoholics.

References

- Clyde, D. J. *Multivariate analysis of variance on large computers*. Miami, Fla.: Clyde Computing Service, 1969.
- Cooley, W. W., & Lohnes, P. R. *Multivariate data analysis*. New York: Wiley, 1971.

- Goldstein, S. G., & Linden, J. D. Multivariate classification of alcoholics by means of the MMPI. *Journal of Abnormal Psychology*, 1969, 74, 661-669.
- Gough, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- Gough, H. G. Appraisal of social maturity by means of the CPI. *Journal of Abnormal Psychology*, 1966, 71, 189-195.
- Gough, H. G., De Vos, G., & Mizushima, K. Japanese validation of the CPI social maturity index. *Psychological Reports*, 1968, 22, 143-146.
- Hogan, R. Development of an empathy scale. *Journal of Consulting and Clinical Psychology*, 1969, 35, 58-63.
- Hogan, R., Mankin, D., Conway, J., & Fox, S. Personality correlates of undergraduate marijuana use. *Journal of Consulting and Clinical Psychology*, 1970, 35, 58-63.
- Huber, N. A., & Danahy, S. Use of the MMPI in predicting completion and evaluating changes in a long-term alcoholism treatment program. *Journal of Studies on Alcohol*, 1975, 36, 1230-1237.
- Jackson, D. N. *Personality Research Form manual*. Goshen, N.Y.: Research Psychologists Press, 1967.
- Jackson, D. N., & Messick, S. *Differential Personality Inventory*. London, Ontario, Canada: Author, 1970.
- Kirk, R. E. *Experimental design: Procedures for the behavioral sciences*. Belmont, Calif.: Brooks/Cole, 1968.
- Lawlis, G. F., & Rubin, S. F. 16 PF study of personality in alcoholics. *Quarterly Journal of Studies on Alcohol*, 1971, 32, 318-327.
- Nerviano, V. J. Common personality patterns among alcoholic males: A multivariate study. *Journal of Consulting and Clinical Psychology*, 1976, 44, 104-110.
- Nerviano, V. J., & Gross, W. F. A multivariate delineation of two alcoholic profile types on the 16 PF. *Journal of Clinical Psychology*, 1973, 29, 371-374.
- Partington, J. T., & Johnson, P. G. Personality types among alcoholics. *Quarterly Journal of Studies on Alcohol*, 1969, 30, 21-34.
- Skinner, H. A., Jackson, D. N., & Hoffman, H. Alcoholic personality types: Identification and correlates. *Journal of Abnormal Psychology*, 1974, 83, 658-666.
- Skinner, H. A., Reed, P. L., & Jackson, D. N. Toward the objective diagnosis of psychopathology: Generalizability of modal personality profiles. *Journal of Consulting and Clinical Psychology*, 1976, 44, 111-117.
- Sutherland, E. H., Schroeder, H. G., & Tordella, A. B. Personality traits and the alcoholic: A critique of existing studies. *Quarterly Journal of Studies on Alcohol*, 1950, 11, 547-561.
- Syme, L. Personality characteristics and the alcoholic: A critique of current studies. *Quarterly Journal of Studies on Alcohol*, 1957, 18, 288-302.
- Tamerin, J. S., & Neumann, C. P. The alcoholic stereotype: Clinical reappraisal and implications for treatment. *American Journal of Psychoanalysis*, 1974, 34, 315-323.
- Whitelock, R., Patrick, J. H., & Overall, J. E. Personality patterns and alcoholic abuse in a state hospital population. *Journal of Abnormal Psychology*, 1971, 78, 1-16.

Received July 18, 1977 ■

Cognitive Preparation and Coping Self-Talk: Anxiety Management During the Stress of Flying

Michel Girodo and Julius Roehl

University of Ottawa and The Royal Ottawa Hospital, Ottawa, Canada

The effectiveness of two cognitive coping strategies, singly and in combination, were investigated in 56 undergraduate females with a reported fear of flying. Subjects were assigned to four groups: preparatory information training, self-statement training, combined (information and self-talk), and pseudotreatment control and were flown aboard an 11-passenger Twin Otter aircraft for two flights. One half of the subjects flew with the door to the cockpit open, the other half flew with the door closed. Each flight encountered a planned unexpected missed landing. Self-reports of anxiety were obtained before takeoff, during the flight, and after landing. Even though the cognitive-coping strategies were not differentially effective in reducing anxiety during the ongoing stress of flying, under serious threat (unexpected event), with the cockpit door open, self-talk and combined subjects coped better than information and control subjects. With the door closed, all groups increased in anxiety. At final landing, with the door closed, self-statement-trained subjects increased in their self-reported anxiety. The results of a 4½-month follow-up on flight apprehension are discussed in light of the effects of the treatment manipulations.

Stress inoculation training (SIT) refers to a cognitive-behavioral treatment package (Meichenbaum, 1975) in which patients are taught to emit positive coping self-statements in learning how to cope with such stressors as pain (Turk, Note 1), test anxiety (Meichenbaum, 1972), and anger problems (Novaco, 1976). Basically, the SIT procedure involves (a) providing the subject with a theoretical framework or rationale for conceptualizing the effectiveness of self-talk in coping with a stress reaction; (b) teaching the subject a series of positive self-statements that focus on (i) preparing for a stressor, (ii) confronting and handling a stressor, (iii) coping with a feeling of being overwhelmed, and (iv) reinforcing oneself for having coped; and (c) trying out or rehearsing coping self-talk.

The authors are grateful for the assistance provided by Air Transit and the University of Ottawa for financial support in conducting the study.

Requests for reprints should be sent to Michel Girodo, School of Psychology, University of Ottawa, 1245 Kilborn Avenue, Ottawa, Ontario, Canada K1H 6K9.

Coping self-talk procedures have also been used in connection with the reduction of psychological stress in surgical patients by Langer, Janis, and Wolfer (1975). In this study, patients who were taught cognitive reappraisal of anxiety-provoking events and were induced to engage in calming self-talk and attention diversion coped better with pre-surgery and postsurgery stresses than did the subjects who were given preparatory information designed to induce "work of worrying" and subsequent emotional inoculation for the stressor (Janis, 1958, 1971). Although the preparatory information increased arousal, this effect dissipated over time, and no evidence of positive postoperative effects were found. These results are in contrast to the findings of Melamed and Siegel (1975), who studied the effects of filmed modeling in reducing anxiety in children facing hospitalization and surgery. The results of this study support the contention that a moderate amount of arousal prior to the stressor may facilitate coping with the stressor once it presents itself.

Another way of looking at the work of worrying process focuses on the notion that the

beneficial effects of engaging in work of worrying are obtained when the person engages in anticipatory problem solving and cognitive rehearsal of coping with the forthcoming stressful event (Meichenbaum, 1975). The suggestion is that the individual experiences a moderate level of arousal when covertly rehearsing the handling of a stressful situation, and this prompts the emotional inoculation necessary to help the person cope better with the stressor.

The general purpose of the present experiment was to assess the applicability of the SIT procedure and preparatory information in coping with a real-life stressor. Specifically, the experiment sought (a) to examine the relative effectiveness of (i) the SIT procedure and (ii) preparatory information, singly and in combination, in coping with a flying experience among persons who were apprehensive of flying; (b) to study the effects of these treatment procedures over two trials; and (c) to examine the effects of an additional stressor in the form of an "unexpected missed landing" on coping ability. It was felt that an unexpected stressful event during the flight might provide the needed variation to demonstrate greater differential effectiveness of one treatment over the other.

An additional feature of the experiment concerned whether subjects had visual access to the pilot and copilot in the cockpit area. In this regard, it was hypothesized that subjects who received preparatory information in the form of reliance on "danger control authorities" would experience less anxiety under conditions in which such figures were visually present and salient, compared to preparatory information subjects who did not have access to such visual information and external reassurances. It was predicted that subjects trained in the SIT procedure would not be as reliant on danger control authorities or on seeking external reassurances compared with individuals given preparatory information, and, as such, that they would be less susceptible to the effects of the availability or nonavailability of these external cues.

Method

Subjects

Subjects were 56 female undergraduate students ranging in age from 18 to 34 ($M = 21$). They were

selected on the basis of four screening questionnaire criteria: (a) They scored 6 or more on a 10 point scale designed to assess flight apprehension (i.e., "How do you usually feel while flying in an airplane?" Very calm and relaxed served as the anchor at one end, and very nervous and tense provided the anchor at the other end). (b) They had previously flown (M no. flights = 5.48). (c) They had never flown in the particular type of aircraft used in the present experiment. (d) They had indicated on the screening questionnaire that they would be willing to serve as a subject in an experiment that involved flying. The screening questionnaire was completed by 1,526 male and female students enrolled at the University of Ottawa, Carleton University, and Algonquin College. Of these, 92 females met the four selection criteria.¹ These subjects were contacted on a random basis and were scheduled for participation on 1 of 6 weekends according to their availability.

Aircraft and Apparatus

The aircraft in which subjects were flown consisted of a DeHavilland Twin Otter with a capacity of two crew and 11 passengers. This short takeoff and landing (STOL) aircraft takes off and lands using a 1,500 foot (457.2 m) runway at a 6-degree angle (compared with the conventional 3-degree slope). In addition to standard-specialized avionics, the STOL aircraft was equipped with a computerized air data acquisition system, which recorded 43 performance parameters of the flight. Of interest to study for possible covariate analysis were measures of turbulence in the form of vertical deviation scores obtained every 2 sec. Earphones and a hand-held microphone were installed at the rear of the cabin to allow for constant communication between the cockpit area and the experimenter. A sliding door served to separate the pilot and passenger area for half of the flights. The seating in the STOL was the same as that found in modern passenger aircraft, with five seats on the left and six seats on the right of the aircraft.

Design

Four training groups of 14 subjects each were exposed to one of four treatments or treatment combinations. One group of subjects was exposed to preparatory information; a second group, to the SIT procedure (self talk); a third group, to a combination of the information and self talk treatments (combined); and a fourth group of subjects was exposed to films on the history of aviation (control).

Subjects were seen in small groups for training on the Saturday preceding the Sunday flight. All subjects participated in two flying trials, the first involving a 45-minute flight from Ottawa to Montreal with a 20-minute stopover, followed by a second identical trial returning to Ottawa. Subjects in the same treat-

¹ The distribution of scores for the 943 females had a mean of 3.0 with a standard deviation of 2.8. The cutting score of 6 ($z = 1.29$) represented the upper 10% of the population.

ment groups were assigned seats behind each other on the same side of the aircraft. This served to reduce possible social comparison processes; persons visible to any one subject were essentially strangers in a different treatment condition. All subjects flew during both trials with either the door open or closed. Measures of self-report of anxiety were obtained at six points throughout the first trial: during "rev up"; after takeoff; following 10 minutes, 20 minutes, and 30 minutes of cruising; and after landing. On Trial 2, an additional measure of self-report of anxiety was obtained following the "missed approach." Thus, this formed the basis of a 4 (groups) $\times 2$ (door) $\times 2$ (trials) $\times 6$ or 7 (assessment periods) factorial design.

On the second trial, the missed landing occurred as follows: In its final approach into Ottawa, the aircraft descended in a normal landing pattern to an altitude of 100 feet (30.4 m). At this point a stall-warning horn sounded in the cockpit, the pitch of the propellers was changed, and full power was applied to the engines. The aircraft rose rapidly, banking to the left until it reached an altitude of 2,000 feet (609.6 m). At this point, the captain apologized for the missed approach, explaining that a light aircraft was taxiing onto the runway without authorization from the control tower and that they would be making a normal landing in about 5 minutes.

Procedure

Instructions and specific training procedures were delivered primarily via audio tape recorder. All subjects were assured that their responses would be kept confidential and that no harm would come to them in the normal course of the experiment. The experimenter also requested that subjects not disclose details of the experiment to anyone until they had received a statement in the mail describing the purpose and results of the experiment and the role they had played. Following this, subjects were asked to complete a self-report of anxiety inventory (SRAI) and the initial flight apprehension inventory (FLAPI) and were asked to sign a consent form.

Self-Talk Training

This 2½-hour training procedure attempted to follow as closely as possible the descriptions given by Meichenbaum (1975) in outlining the SIT procedure. Briefly, as part of this training, the experimenter stressed the importance of learning positive self-statements as a method for coping with stress. All subjects were asked to share with other members in the group events in which they experienced nervousness or anxiety and to recall the negative thoughts that accompanied these situations. Through a tape recording, the conceptual rationale of the SIT and a list of positive self-statements in the four phases of coping with a stressor were presented. Taking a slow deep breath was presented as a effective strategy for dealing with the physical component of anxiety, and positive coping self talk was presented as the strategy for handling the cognitive component of anxiety. The tape continued with an example of a person taking an exam in which negative

self-statements and feelings of anxiety were present, followed by instructions that negative self-statements and anxiety feelings should serve as cues for emitting positive-coping self-statements. Following this, subjects were given a prepared list of coping self-statements (Meichenbaum, 1975). They were then instructed to imagine a stressful situation, to prepare their own list of coping self-statements, and to memorize this list. Following a 10-minute learning period, subjects were tested on their recall of the self-statements. To further consolidate the training procedure, the tape continued with a description of an imaginary horseback ride for a person who was fearful of horses. Subjects were asked to fantasize and imagine clearly a person making positive coping self-statements as the horse was approached, touched, mounted, ridden, and dismounted after a successful riding experience. Following this exercise, subjects were asked to share with other members of the group their individual self-statements, and each subject compiled the final list of positive self-statements that she was to use when flying the next day. Subjects were instructed to rehearse making positive coping self-statements in anticipation of the flight of the following day.

Preparatory Information Training

As in the self-talk training, this procedure began with the same request for a discussion of events in which subjects experienced nervousness or anxiety. Following this, the experimenter expressed the importance of acquiring prior information about potentially stressful events as a method for coping with stress. The experimenter then illustrated the effectiveness of prior information for coping with test anxiety. A tape recording then described the theoretical rationale of this training procedure, drawing examples from the scientific literature (e.g., Janis, 1958, 1971). The audiotape continued with a detailed description of the events that were to take place the following day. This description included the nature of the transportation to the airport, the preboarding procedures, the exterior and interior of the aircraft, and the takeoff and landing procedures. Twelve colored slides depicting the STOLport, aircraft, a view from the ground over the two cities and during cruising, takeoff, and landing were shown at appropriate points in the presentation. To have the subjects view the pilots as "danger control authorities," descriptions of the pilots' training and experience with emphasis on their skill, expertise, maturity, and 20 years of service were given. To control for time with the experimenter, the above presentation was given twice.

Combined Training

This training procedure consisted of a 3½-hour combination of the SIT and preparatory information procedures. To control for equal time with the experimenter, the information procedure was presented once.

Pseudotreatment Control

In this condition, as in the other three treatments, subjects spent an initial period of time sharing with the group how they felt and what they thought of when in a stressful situation. This was designed to control for possible effects associated with ventilation of affect related to fears. Following this, subjects were shown three 20-minute films depicting the history of aviation in Canada. Following the film presentations, subjects were asked to relate to the group what they liked least and most about each film. Thus, time with the experimenter and interaction time among subjects was equivalent across all conditions.

To reduce time spent completing scales on the day of the flight and after training, subjects in all groups were given specific instructions and practice with the SRAI that they would be using the following day. Subjects were instructed not to talk to each other at any time at the STOLport or throughout the flight.

Dependent Measures²

The SRAI served as the chief dependent measure in the experiment. This inventory consisted of four 11-point scales anchored at each end from very to not at all, and was designed to measure cognitive and somatic components of anxiety after the symptom clusters of Buss (1966). The four scales measured (a) calm and relaxed; (b) worried and apprehensive; (c) anxious and nervous; and (d) tense and trembling. Item analyses correlating each scale score with total scale score revealed that essentially the same construct was being tapped, consistent with coefficients obtained in previous studies (Girodo, 1974).

The SRAI was administered prior to training, immediately following training, at the training site, the morning of the flight at the STOLport prior to boarding, during the rev up, following takeoff, after three 10-minute cruising periods, following the unexpected event on Trial 2, and after landing. The SRAI was modified slightly on several of these occasions to coincide with the events that had just transpired. Thus, although subjects were generally asked to indicate how they felt at present, the estimates obtained for the three 10 minute cruising periods asked the subjects to indicate how they had felt during the last 10 minutes or how they had felt as a result of the takeoff, the unexpected event (see below), and the landing. Another modification asked, "How have you been feeling since yesterday afternoon in anticipation of the flight today?" Subjects used the SRAI scales to reply to this item when they met at the training site on the morning of the flight. Following the missed approach, the experimenter stated that he wanted to take advantage of the situation and record reactions on an additional questionnaire. This he did by distributing an untitled SRAI and verbally requested the subjects to indicate on the scales provided how they had felt as a result of what had just happened.

Treatment effectiveness questionnaire. This consisted of an eight-item postexperimental questionnaire designed to measure the effectiveness of self-statements in coping with the stress of flying. Here, self-talk and

combined group subjects were asked to (a) list the self-statements that they emitted during the flight; (b) indicate the percentage of time that they used self-talk, and (c) rate the effectiveness of self-talk as a coping device on an 11-point scale ranging from not as helpful and useful at one end to very helpful and useful at the other end. Subjects in the information condition were asked to (a) rate the effectiveness of the information given to them on a similar 11-point scale in helping them cope with the stress and (b) indicate what kind of information might have been beneficial to them in coping with the flight.

Follow-up

A follow-up letter and questionnaire were mailed to all subjects 4½ months after the conclusion of the experiment. Although this was designed to assess any negative or untoward effects of participating in the experiment, it also inquired into the current flight apprehension of subjects by repeating the FLAPI, given at the time of initial screening.

Results

Training and Preflight Assessment

An analysis of variance performed on SRAI scores obtained immediately after training indicated that treatment manipulations produced significant differences in self-report of anxiety between groups, $F(3, 52) = 8.83$, $p < .001$. Tukey tests indicated that both self-talk and combined subjects reported significantly more anxiety ($p < .05$) immediately after training compared with subjects in the control and information conditions. The scores obtained on the SRAI the morning of the flight at the training site failed to yield significant group differences; however, the results obtained by using the Duncan's multiple-range test (Duncan, 1955) suggested that self-talk subjects were still slightly more aroused than information subjects ($p < .05$). A similar analysis of SRAI scores obtained at the STOLport prior to departure yielded a similar non-significant F ; however, again Duncan's test revealed that self-talk subjects were slightly more aroused than subjects in the combined group ($p < .05$).

² At the STOLport, electrodes were attached to the subjects' wrists and left ankle to record heart rate on cassette. Equipment problems early in the experiment forced the experimenter to abandon this measure; however, electrodes were still applied to all subjects on the remaining flights.

Table 1

Means and Standard Deviations of Self-Report of Anxiety Inventory (SRAI) Scores at Pretraining and Posttraining and Work of Worrying Scores on Flight Day

Group	SRAI				Work of worrying scores	
	Pretraining		Posttraining		<i>M</i>	<i>SD</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Information	11.65	7.67	6.43*	4.18	8.60	4.86
Self-talk	14.29	6.93	14.36	6.42	15.50**	5.00
Combined	17.21	7.81	12.29	7.26	8.14	5.33
Control	13.43	7.71	5.29*	3.52	8.29	5.20

* $p < .05$; significant decreases from pretraining to posttraining.

** $p < .001$; significantly greater than mean scores from other groups.

An analysis of variance performed on the flight anticipation scale that asked how subjects had been feeling since yesterday afternoon in anticipation of the flight showed a significant treatment group effect $F(3, 52) = 6.96$, $p < .001$. Tukey tests indicated that self-talk subjects reported engaging in significantly more work of worrying than subjects in the other three groups. Table 1 summarizes the results of the SRAI and "worry" scales for these preflight assessments.

Flight Assessment

It was found that although turbulence did take place and was perceptible by the pilots, it had no significant effect on SRAI reports.³

Trial 1. A 4 (groups) $\times 2$ (door) $\times 6$ (assessment periods) analysis of variance on the SRAI scores yielded a significant effect for assessment periods, $F(5, 48) = 35.96$, $p < .001$. Treatment groups did not differ from each other at either rev up, takeoff, following each of the three 10-minute cruising periods, or during landing. An analysis of variance on the Groups \times Door \times Cruising Periods produced a main effect for cruising periods, $F(2, 48) = 8.31$, $p < .001$. A main effect for the door condition was obtained for the third cruising period, $F(1, 48) = 4.73$, $p < .03$, and Tukey tests revealed that subjects who flew with the cockpit door open reported significantly more arousal ($p < .01$) than subjects who flew with the door closed. No SRAI differences were found between groups for the door condition for the first landing.

Trial 2. A $4 \times 2 \times 7$ analysis of variance on SRAI scores produced a main effect for the door condition, $F(1, 48) = 5.48$, $p < .03$, and a main effect for assessment periods, $F(5, 48) = 47.69$, $p < .001$. An analysis of variance considering only the treatment groups and the rev up and takeoff periods failed to produce significant differences. An analysis of variance on the groups, door condition, and three cruising periods yielded a significant effect for door, $F(1, 48) = 8.90$, $p < .004$, and for cruising periods, $F(2, 48) = 11.46$, $p < .001$. Even though treatment groups did not differ from one another at any of the cruising periods, subjects in the door open or door closed conditions differed significantly from each other after the first 10 minutes, $F(1, 48) = 10.32$, $p < .003$; following 20 minutes of cruising, $F(1, 48) = 5.90$, $p < .02$; and following 30 minutes of cruising, $F(1, 48) = 8.20$, $p < .006$. Figure 1 illustrates the nature of these findings.

Subjects who flew with the door open reported significantly more arousal than subjects who flew with the door closed when confronted by the unexpected event, $F(1, 48) = 4.25$, $p < .04$. At final landing, however, no significant overall differences were found between door open or door closed conditions.

Since the door conditions produced significant differences in SRAI scores, separate post hoc analyses were performed on the third

³ A complete report on the way the air data acquisition system and crew reports of turbulence were treated in analyses of covariance is available from the first author.

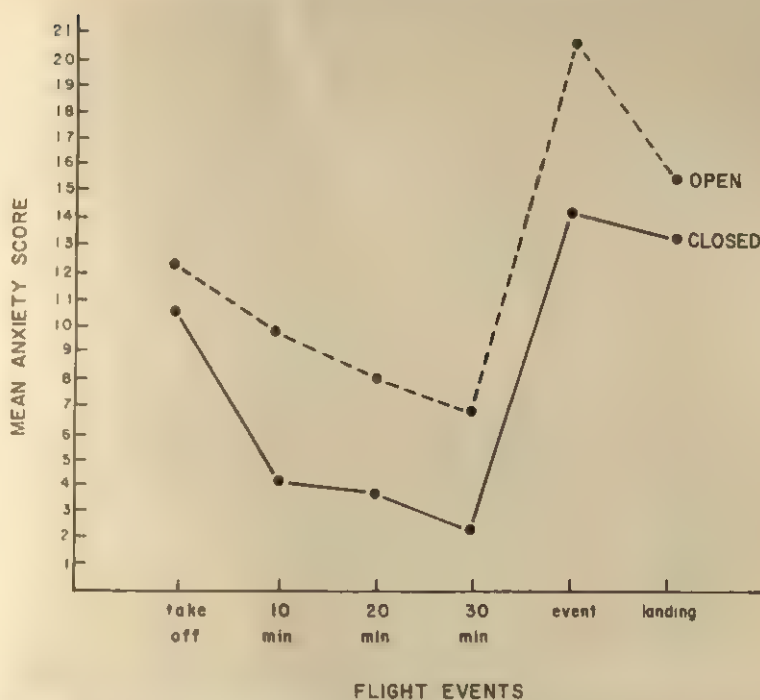


Figure 1. Mean self-report of anxiety inventory scores for subjects in the door open and door closed conditions across flight events for the second trial.

cruising period, the unexpected event, and the final landing. Figure 2 plots the SRAI scores for subjects in the treatment groups across these three periods for door open and door closed conditions.

When the door open condition is examined, it can be seen that significant differences between self-talk subjects and control subjects were obtained during the unexpected event, $t(12) = 3.10$, $p < .009$. Again, in the door open condition, the changes in SRAI scores from the last cruising period to the unexpected event were examined by correlated t tests. It was found that the SRAI scores of subjects in the control condition and in the information condition significantly increased from the last cruising period to the unexpected event, $t(6) = 4.99$, $p < .02$, and $t(6) = 3.39$, $p < .002$, respectively. Even though there was a tendency for the scores of both self-talk and combined group subjects to increase, this was not statistically significant. From the unexpected event to the landing, all groups tended to show decreases in their self-report of anxiety; however, only the control subjects' scores de-

creased significantly from the unexpected event to the landing, $t(6) = 2.53$, $p < .05$.

In the door closed conditions, the scores of subjects in all four treatment groups increased significantly but not differentially. Considering the SRAI scores from unexpected event to landing, correlated t tests revealed that although the SRAI scores of the control group subjects decreased significantly, $t(6) = 2.48$, $p < .05$, together with a nonsignificant tendency for the scores of the combined group subjects to decrease, the scores of self-talk subjects actually *increased* from unexpected event to landing, $t(6) = 4.58$, $p < .004$. Information subjects showed a nonsignificant increase in self-report of anxiety.

In view of the wide variability of the data at each of the plotted points, complex chi-square analyses were performed on the number of subjects whose scores decreased as opposed to those whose scores stayed the same or increased. From the last cruising period to the unexpected event, scores for *all* subjects except one (self-talk, door open) increased, $\chi^2(3) = 3.06$, ns ; however, differential changes in SRAI

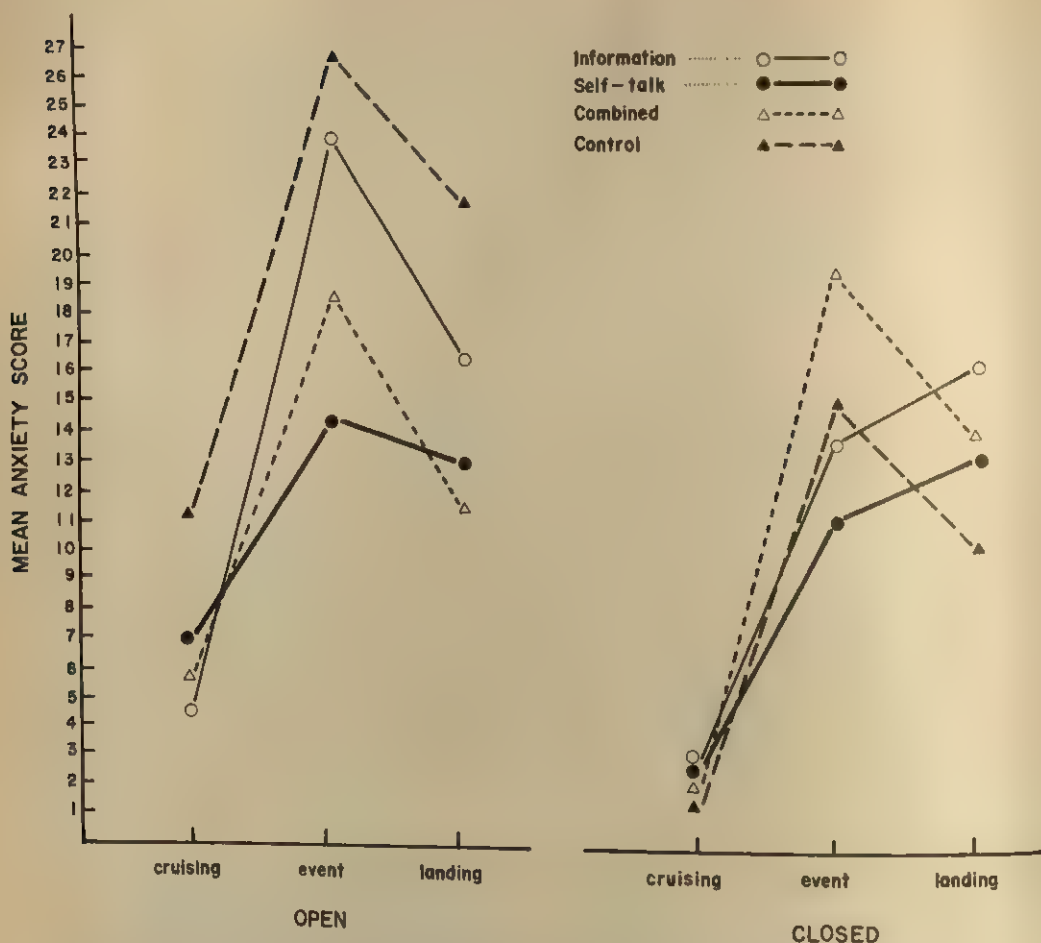


Figure 2. Mean self-report of anxiety inventory scores for subjects in the four treatment groups, in door open or closed conditions, at the last cruising period, the unexpected event, and the final landing. (From *Stress and Anxiety* (Vol. 4) by I. G. Sarason and C. D. Spielberger (Eds.), 1977. Copyright 1977 by Hemisphere. Reprinted by permission.)

scores from unexpected event to landing were obtained, $\chi^2(3) = 18.89$, $p < .005$, in that scores of 11 of the 14 subjects in the control group and 11 of the 14 subjects in the combined group decreased from unexpected event to landing, whereas only 3 of the 14 self-talk subjects and 7 of the 14 information subjects obtained lower SRAI scores from unexpected event to landing. When the door conditions were examined for chi-square differences, only the door closed chi-square obtained differences in number of subjects who either increased or decreased from unexpected event to landing, $\chi^2(3) = 15.43$, $p < .02$. Here, 5 of 7 control subjects and 5 of 7 combined subjects showed decreases in anxiety, whereas only 3 of 7 in-

formation subjects and none of the self-talk subjects reported decreases in anxiety. Fisher's exact test (Fisher, 1966) revealed that there were significantly more control (5) and combined (5) subjects' SRAI scores that decreased compared with none of the subjects in the self-talk group, $\chi^2(3) = 7.69$, $p < .05$.

Seating. An analysis was undertaken for subjects seated in the first four seats nearest the cockpit and the remaining six seats to the rear of the aircraft. On Trial 1, a Group \times Door \times Seating analysis of variance failed to reveal any significant effect for seating or any interaction between seating and the door or group condition. Similar nonsignificant effects for seating were obtained on SRAI scores for

Trial 2. However, a significant Group \times Seating interaction was obtained for the landing on Trial 2, $F(3, 40) = 4.39$, $p < .009$. Tukey tests showed that at the final landing, information group subjects who sat in the front of the aircraft reported the lowest SRAI scores, whereas information subjects seated in the rear reported the highest SRAI scores. Subjects in the combined group and in the self-talk group who were seated in front had higher SRAI scores than subjects in those groups who sat in the rear of the aircraft. Control group subjects did not differ significantly in terms of their assigned seating. It was interesting to find that these effects were only slightly amplified (but not significantly) in the door open as opposed to the door closed condition.

Postexperimental Questionnaire

In reply to the question "Do you wish you had acquired more information and known better what to anticipate today?", seven of 14 control subjects expressed the need for more information, compared with 2 of 14 information subjects and 2 of 14 combined subjects, $\chi^2(2) = 6.16$, $p < .05$. Fisher's exact test revealed that more subjects in the information group (12) expressed satisfaction with the amount of information that they had acquired prior to the flight compared with subjects in the control group (7) ($p < .05$). When subjects were asked to rate how effective they had found the techniques of providing information about what to anticipate in helping them cope with the stress of flying, no differences were found between subjects in the door closed condition. However, subjects in the door open condition differed significantly in their ratings on this question, $F(2, 18) = 12.12$, $p < .001$. Tukey tests revealed that information and combined subjects did not differ in their rated effectiveness of acquiring information as a coping device. However, control subjects rated the information that they had acquired as significantly less effective in coping with stress compared with information or combined subjects ($p < .05$). The self-talk and combined subjects, overall and in both door conditions, did not differ on the percentage of time that they had spent emitting positive self-statements. The subjects' ratings on the

effectiveness of emitting positive self-statements in coping with the stress of flying also revealed no differences between these two groups. A significant positive correlation ($.61$, $p < .001$) was obtained between the reported percentage of time that subjects spent making positive self-statements and the rated effectiveness of their making self-statements in coping with the flight experience. Also, positive correlations were obtained between SRAI scores at various assessment periods and the reported percentage of time spent making positive self-statements. Specifically, positive correlations between anxiety scores and self-statement use were obtained at the third cruising period for Trial 1 ($r = .34$, $p < .04$), at the first landing ($r = .32$, $p < .05$), and for the unexpected event ($r = .33$, $p < .04$). No significant correlations were obtained between SRAI scores at each of the assessment periods and the reported effectiveness of self-talk.

Follow-up

Fifty-three of 56 (95%) follow-up questionnaires mailed 4½ months after the study was completed were returned. Subjects in all four treatment groups reported significantly lower flight apprehension scores at follow-up compared with their pretraining scores. For subjects who were in the door open condition, the information, self-talk, and combined groups, $t_s(6) = 3.58$, 4.46, and 4.70 ($p < .01$), respectively, reported significant decreases in their FLAPI scores, whereas control group subjects failed to show any significant change in their FLAPI scores from pretraining to follow-up. For subjects in the door closed condition, t tests on change scores revealed significant decreases in flight apprehension for self-talk subjects, $t(6) = 6.30$, $p < .001$; and combined subjects, $t(6) = 3.12$, $p < .03$. Subjects in the information and control conditions did *not* report significant changes in FLAPI scores. Thus, subjects who were exposed to self-statement training (self-talk and combined) reported significant reductions in flight apprehension 4½ months after the experiment regardless of the condition that they were in. It is of interest to note that subjects in the control group failed to show significant decrements in flight apprehension from pretraining to follow-up, especially in view of the

fact that their SRAI scores were markedly similar to those of the treatment groups at various points throughout the two trials.

Discussion

The major findings of the experiment were as follows: (a) Generally, subjects who flew with the door open showed more anxiety than subjects who flew with the door closed. Even though all subjects tended to decrease in anxiety in the form of possible habituation effects (e.g., Solyom, Shugar, Bryntwick, & Solyom, 1973), door closed subjects habituated more rapidly than door open subjects. (b) The various treatment manipulations generally failed to demonstrate differential coping effectiveness throughout the normal course of the flights. (c) The additional stressor, in the form of an unexpected event, produced differential increases in anxiety when the cockpit door was open, in that self-statement-trained subjects coped better than information and control subjects. This was not the case in the door closed conditions in which self-talk subjects did not cope with the subsequent landing as well as subjects in other groups.

Work of Worrying

No evidence was found to support the hypothesis that preparatory information served to increase arousal levels either immediately after training or prior to the flight. Only SIT-trained subjects reported significantly high levels of anxiety immediately following training as well as prior to the flight itself. Also, self-talk subjects also reported greater amounts of "worry and apprehension" during the 20 hours preceding the flight. It is interesting to note that combined subjects who also received self-statement training did not manifest the same increase in arousal following, nor did they report greater work of worrying, prior to the flight. Conceivably, self-talk subjects may have rehearsed coping self-statements, in connection with imaginary and nonspecific flight events, unlike the combined subjects who may have rehearsed their self-talk with predictable events and images depicted on the slides. Whatever the effects of satisfactory worry/cognitive preparation on coping with a

future stressor may really be, they were not obtained in the present study: Recall that both self-talk and combined subjects coped comparatively well with the unexpected event in spite of the fact that combined and information subjects failed to reveal any indication of anticipatory anxiety.

Door Open Condition

As might be expected, preparatory information seemed to be ineffective in coping with an unexpected missed landing, and subjects in this condition responded to stress with increases in anxiety similar to the control subjects. Also, even when danger control authorities were visible, information subjects did not appear to find their visible presence a source of reassurance to allay their anxieties.

Door Closed Condition

Here, it appears as if not having visual contact with the cockpit area can account for the reasons that the treatment groups significantly increased in anxiety from the last cruising period to the unexpected event. It appears that in the door closed condition, the importance of being able to process information about the significance of a stressor may take precedence over any attempted coping strategy. Conceivably, being able to monitor what is happening or to be able to answer questions one is raising concerning the significance and importance of an unanticipated event may have an overriding influence, and no amount of self-talk will serve to distract the individual from this information search process.

The effects of seating arrangement appeared to be most salient for the information group. Subjects in the information group seated in the front of the aircraft reported the lowest SRAI scores of all treatment groups, and the information subjects seated at the rear obtained the highest scores. At first glance it would appear as if subjects exposed to the information coping strategy were better able to cope with the second landing if they were able to process information, confirm expectations, and rely on cues from danger control authorities for obtaining reassurance. However, this seating effect on information subjects was obtained in *both* the door open and

door closed conditions. Although this is somewhat puzzling, it could suggest that physical proximity from such danger control authorities is a more important variable than the availability of visual contact with such figures when it comes time to cope with such a stressor. It is interesting to note that since the effect of proximity to the cockpit area was not obtained for subjects given other coping strategies, subjects given preparatory information could have been excessively reliant on this coping strategy in coping with the stress of the second landing.

The seating factor only became significant for the information subjects during the second landing. This is understandable if we consider that it is only after the missed landing that these subjects felt an additional stress in the sense of a disconfirmation of the expectancies induced by the preparatory information treatment package. Possibly because they could no longer rely on the information presented to them concerning what to anticipate, these subjects may only at that time have felt a need for reassurances or for proximity to danger control authorities.

The follow-up data suggest that the SIT and preparatory information procedures singly or in combination were equally effective in reducing flight apprehension 4½ months later. This was true only for subjects who flew with the cockpit door open, since only subjects trained in the SIT procedure who flew with the door closed reported long-term decrements in flight apprehension. The fact that door closed/information group subjects did not report lowered flight apprehension at follow-up may relate to an assertion by Averill (1973) in connection with perceived control over aversive events. He suggested that even though information about a stressor may have value in helping a person cope with a stress, it may only be effective if it is validated by experience and if such information was found to be veridical in reducing objective worry following reality testing. It may not be too surprising, therefore, to have found that preparatory information was not useful in producing long-term decrements in flight apprehension especially when access to sources of reassurance was cut off and after expectations concerning the flight events were disconfirmed.

Theoretical Considerations

Two important theoretical issues are raised by the results of the present investigation. First, looking only at the anxiety scores obtained during the two cruising periods, the absence of any evidence of a differential coping ability across treatment groups suggests that the SIT procedure is no more effective for coping with the ongoing stress of flying than preparatory information or when persons are left to their resources (i.e., control). These findings cast some doubt on the theoretical underpinnings of the SIT procedure, specifically on the assertion that the act of emitting positive self-statements in the presence of a stressor can effectuate anxiety reduction.

The second theoretical issue concerns the mediating role of cognitions in altering emotional reactions. Since, in the present experiment, self-talk/door open subjects did not increase in anxiety during the unexpected event, and since the anxiety scores of all seven of the self-talk/door closed subjects increased from the unexpected event to the final landing, the issue does merit some attention. We can focus on the specific question, "What cognitive processes can explain the mediating role of self-talk in stress reactions?" First, it might be useful to determine what processes are probably not involved before trying to arrive at an understanding of what they might be. If we look at the semantic therapies of Beck (1970) and Ellis (1962), for example, a common denominator of cognitive reorganization serves as the foundation for coping and emotional adjustment. Whether it be the correction of arbitrary inferences or selective abstractions in the former, or the changing of irrational beliefs and assumptions in the latter, both cognitive therapies work at the level of the person's epistemological base for creating new perceptions and understandings. As has been explained elsewhere (Girodo, 1977), in spite of certain similarities, the crucial cognitive processes in the SIT procedure may have very little to do with modifying the subject's epistemological base. When subjects "buy" the persuasive conceptual rationale of the SIT package, this does not necessarily cast the die for producing emotional responses simply by repeating their semantic equiva-

lents. Indeed, as was found in a coping with pain study by Girodo and Wood (Note 3), the persuasive rationale serves more to motivate subjects to use the self-talk technique rather than to reorganize cognitions that would make the stressor less painful. Inasmuch as attentional processes were implicated in producing differential coping responses in the SIT pain experiment, we felt that it might be more than simple conjecture to suggest that similar attentional processes might be involved in SIT subjects' responses to the unexpected event and subsequent landing.

What is it about monitoring internal states and repeating self-statements that had such a detrimental effect on SIT door closed subjects during the final landing? Although it is conceivable that these subjects abandoned their self-talk strategy after the unexpected event (possibly because it may have failed them when it was most needed), an alternative explanation along the lines proffered by Sarason (1975) is more compelling. He suggested that self-preoccupation involves the kind of attentional activity that can interfere at a variety of points with information processing and subsequent planning strategies for coping with a stressor. The conditions for increased self-preoccupations were present in the self-talk treatment in that subjects were exposed to (a) a door closed condition in which cockpit and pilot reassurance cues were not available and (b) an indoctrination to a self-statement strategy that forced subjects to attend mainly to their internal state. We propose that this kind of induced self-preoccupation may have interfered with information processing and evaluation necessary in the planning of a coping strategy for dealing with the uncertainty of the second landing. Why this increase in anxiety at the final landing was not manifested in combined subjects is difficult to explain unless we assume that the reassurances concerning pilot competency had an appropriate effect then.

Finally, recall that significant positive correlations were obtained between the reported percentage of time that subjects were making positive self-statements and their anxiety scores at three of the assessment periods. In light of (a) the previous arguments suggesting that forced self-talk may interfere with appraisal and coping processes, (b) the finding

that two of these correlations were obtained for stressors that would normally invite information processing and/or appraisal processes (i.e., the first landing and the unexpected event), and (c) the absence of a significant correlation between the subjects' rated effectiveness of the SIT procedure and their anxiety scores, it is reasonable to suggest that the more time subjects spent making their self-statements, the more anxious they became at various points of the flying experience.

In conclusion, we would like to draw attention to the distinction between semantic therapies that work on inducing new beliefs or correcting cognitive distortions, and thus produce emotional change via this epistemological base, and the self-talk therapy in which the mediating processes may be more inimical. We should not lose sight of the fact that the conceptual rationale underlying the SIT procedure is designed to induce the subject to comply with self-talk instructions and that they are given for *his or her* benefit, and not for the therapist to come to believe. On the basis of the present experiment, we suggest that if indeed attention information-processing demands mediate in SIT-produced emotional reactions, then these can be viewed as methodological artifacts of the semantic therapy. On the one hand, maybe we better not let patients in on this, for it may be only because they believe that their positive self-talk can produce their semantic equivalent that they cope better on occasion. On the other hand, we should not forget that the self-talk may be applied so faithfully that it interferes with the use of other available coping mechanisms.

Reference Notes

1. Turk, D. *Cognitive control of pain: A skills training approach*. Unpublished manuscript, University of Waterloo, Waterloo, Ontario, Canada, 1974.
2. Meichenbaum, D., & Turk, D. *The cognitive-behavioral management of anxiety, anger, and pain*. Paper presented at the Seventh Banff International Conference on Behavior Modification, Alberta, Canada, March 1975.
3. Girodo, M., & Wood, D. *Talking yourself out of pain: The relative effects of believing in a coping strategy*. Unpublished manuscript, University of Ottawa, Ottawa, Ontario, Canada, 1978.

References

- Averill, J. R. Personal control over aversive stimuli and its relationship to stress. *Psychological Bulletin*, 1973, 80, 286-303.
- Beck, A. Cognitive therapy: Nature and relation to behavior therapy. *Behavior Therapy*, 1970, 1, 184-200.
- Buss, A. H. *Psychopathology*. New York: Wiley, 1966.
- Duncan, D. B. Multiple range and multiple *F* tests. *Biometrics*, 1955, 11, 1-42.
- Ellis, A. *Reason and emotion in psychotherapy*. New York: Lyle Stuart Press, 1962.
- Fisher, R. A. *The design of experiments*. New York: Hafner, 1966.
- Girodo, M. Yoga meditation and flooding in the treatment of anxiety neurosis. *Journal of Behavior Therapy and Experimental Psychiatry*, 1974, 5, 157-160.
- Girodo, M. Self-talk: Mechanisms in anxiety and stress management. In C. Spielberger & I. G. Sarason (Eds.), *Stress and anxiety* (Vol. 4). Washington, D.C.: Hemisphere, 1977.
- Janis, I. L. *Psychological stress*. New York: Wiley, 1958.
- Janis, I. L. *Stress and frustration*. New York: Harcourt, Brace, & Jovanovich, 1971.
- Langer, E. J., Janis, I. L., & Wolfer, J. A. Reduction of psychological stress in surgical patients. *Journal of Experimental Social Psychology*, 1975, 11, 155-165.
- Meichenbaum, D. Cognitive modification of test anxious college students. *Journal of Consulting and Clinical Psychology*, 1972, 39, 370-380.
- Meichenbaum, D. Self-instructional methods. In F. H. Kanfer & A. P. Goldstein (Eds.), *Helping people change*. New York: Pergamon Press, 1975.
- Melamed, D. G., & Siegel, L. J. Reduction of anxiety in children facing hospitalization and surgery by use of filmed modeling. *Journal of Consulting and Clinical Psychology*, 1975, 43, 511-521.
- Novaco, R. W. Treatment of chronic anger through cognitive and relaxation controls. *Journal of Consulting and Clinical Psychology*, 1976, 44, 681.
- Sarason, I. G. Anxiety and self-preoccupation. In I. G. Sarason & C. D. Spielberger (Eds.), *Stress and anxiety* (Vol. 2). Washington, D. C.: Hemisphere, 1975.
- Solyom, L., Shugar, R., Bryntwick, S., & Solyom, C. Treatment of fear of flying. *American Journal of Psychiatry*, 1973, 130, 423-427.

Received July 20, 1977 ■

Use of Electromyographic Biofeedback and Cue-Controlled Relaxation in the Treatment of Test Anxiety

D. Kenneth Counts, James G. Hollandsworth, Jr., and John D. Alcorn
University of Southern Mississippi

The effect of using electromyographic (EMG) biofeedback to increase the efficacy of cue-controlled relaxation training in the treatment of test anxiety was studied. Forty college undergraduates scoring in the upper third on a self-report measure of test anxiety were randomly assigned to one of four treatment conditions—EMG-assisted cue-controlled relaxation, cue-controlled relaxation alone, attention-placebo relaxation, and no-treatment control. Pre-post self-report measures of test anxiety, state anxiety, and trait anxiety were obtained. In addition, a performance measure in the form of a mental abilities test was administered. Subjects from the three relaxation groups received six 45-minute individual sessions over a period of 2 weeks. All treatments were conducted using audiotape recordings. The results indicate that cue-controlled relaxation is effective in increasing test performance for test anxious subjects, that EMG biofeedback does not contribute to the effectiveness of this procedure, and that self-report measures of anxiety are susceptible to a placebo effect.

The use of relaxation training in the treatment of test anxiety has met with mixed results (Chang-Liang & Denney, 1976; Johnson & Sechrest, 1968). Of primary concern has been the generalization of the relaxation response to stressful situations occurring beyond the confines of the treatment room itself (e.g., Goldfried & Trier, 1974). Modifications of the basic technique, such as cue-controlled relaxation, appear to facilitate this generalization process (Russell & Sipich, 1973). Furthermore, the use of other aids, such as electromyographic (EMG) biofeedback, may increase the effectiveness of relaxation training itself (Haynes, Moseley, & McGowan, 1975). Until now, however, there has been no investigation of the relative contribution of each of these procedures toward improving the test performance of test anxious students.

The technique of cue-controlled relaxation

was first presented as a possible treatment for test anxiety by Russell and Sipich (1973). This approach consists of deep muscle relaxation training and the pairing of a cue word such as "relax" or "calm" with breath exhalation while relaxed. According to the classical conditioning paradigm, after a number of these pairings, the cue word alone should elicit relaxation and a feeling of calmness. Two case studies (Russell & Sipich, 1973, 1974) have provided support for the use of cue-controlled relaxation in treating test anxiety. Russell, Miller, and June (1974) also used group cue-controlled relaxation in the treatment of test anxious college students. Even though the results were encouraging, no control group was included and the dependent measure was restricted to self-reports. A comparison of cue-controlled relaxation with systematic desensitization (Russell, Miller, & June, 1975) indicated that the two treatments were equally effective. Both treatment conditions were superior to the control group on self-report measures, but no differences were noted on the performance measure. Another comparison of these treatment conditions obtained equivocal results (Russell, Wise, & Stratoudakis, 1976). Marchetti, McGlynn, and

The authors would like to thank Robert E. Agnew for his valuable assistance on preparing the treatment tapes.

Requests for reprints should be sent to James G. Hollandsworth, Jr., Department of Counseling Psychology, University of Southern Mississippi, Box 272 Southern Station, Hattiesburg, Mississippi 39401.

Patterson (1977), however, found that the effects of cue-controlled relaxation failed to exceed those of a placebo or no-treatment condition on self-report measures and psychophysiological indices of arousal during test taking.

In recent years, EMG biofeedback training also has gained popularity as a relaxation training method. Essentially, the client receives feedback (auditory, visual, or both) on the amount of tension in the monitored muscle group. This feedback facilitates the relaxation process. Haynes et al. (1975) compared the effectiveness of frontalis EMG biofeedback and two types of verbal relaxation instructions in reducing muscular tension. EMG biofeedback was equal to one form and superior to the other form of verbal instructions. Similar results were found by Reinking and Kohl (1975). Canter, Kondo, and Knott (1975) found EMG biofeedback superior to verbal instructions in reducing tension and concomitantly relieving anxiety symptoms in adult psychiatric patients.

The particular sensory mode of the feedback may be important in EMG relaxation training. Alexander, French, and Goodman (1975) compared the efficacy of auditory and visual feedback in EMG biofeedback training. The results indicated that auditory feedback may be more effective in the induction of muscular relaxation than visual feedback.

The use of EMG biofeedback in conjunction with various behavioral treatment strategies has yielded encouraging results. In one study (Wickramasekera, 1972), EMG biofeedback was used in the relaxation phase of the systematic desensitization of test anxiety. The procedure was supported by decreases in client reports of test anxiety. Reeves and Mealiea (1975) used EMG-assisted relaxation in the cue-controlled relaxation treatment of flight phobia. Even though the Reeves and Mealiea study was without adequate controls, the results indicate that this method has considerable promise as a treatment strategy for a wide range of disorders.

At this point, the majority of research on cue-controlled relaxation with test anxiety has either lacked performance measures, control conditions, attention-placebo conditions, or a combination of these factors. Furthermore, a

controlled investigation of the combination of EMG biofeedback with cue-controlled relaxation training has not been reported.

This study was designed to investigate the effectiveness of EMG-assisted cue-controlled relaxation in the treatment of test anxiety. More specifically, it was hypothesized that test anxious subjects receiving cue-controlled relaxation training or EMG-assisted cue-controlled relaxation training would report less test anxiety and demonstrate a greater increase in test performance than subjects receiving an attention-placebo treatment or no treatment. Furthermore, it was hypothesized that test anxious subjects receiving EMG-assisted cue-controlled relaxation training would report less test anxiety and would demonstrate a greater increase in test performance than subjects receiving the cue-controlled relaxation procedure alone.

Method

Subjects

The Test Anxiety Scale (TAS) was administered to 294 undergraduate students midway through the academic quarter. These students were unaware that high scorers would be given the opportunity to receive treatment for test anxiety. This sample generated a TAS mean score of 16.85 with a standard deviation of 7.97. Sixty-nine students who received a raw score of 21 or above, which placed them in the upper third of the distribution, were contacted by telephone and were asked to participate in the study. A total of 47 subjects were pretested and stratified according to the TAS and Otis-Lennon performance test results. Subjects within each stratum were randomly assigned to one of four treatment conditions—EMG-assisted cue-controlled relaxation, cue-controlled relaxation, attention-placebo, or no-treatment control. Seven of these 47 subjects failed to complete treatment, with the two treatment groups losing 2 each and the attention-placebo group losing 3. As a result, the final sample consisted of 40 subjects, 10 in each group, with a mean age of 20.6 years, ranging from 16 to 35. There were 28 females and 12 males. Most of the subjects were either freshmen or sophomores. Subjects received academic credit for participating in the study.

Apparatus

Subjects received treatments in a dimly lit, air-conditioned, sound-attenuated room in which there was a reclining chair, cassette tape recorder, EMG biofeedback device, small table, and lamp. The room was fitted with a one-way mirror. A Bio-Dyne MR-260 electromyographic biofeedback device providing con-

tinuous auditory (fluctuations in tone) feedback was used in this study. With the exception of two subjects, the sensitivity-gain adjustment was set at 25 μ V. For two subjects the sensitivity-gain adjustment was set at 10 μ V for the initial session only. Three electrodes, each with a surface area of approximately 5.65 cm², were secured with a rubber headstrap approximately 28 cm apart. The electrodes were centered and placed horizontally on the subject's forehead. The frontalis surface area was prepared with isopropyl alcohol. Cor-Gel electrocardiogram electrode gel was used as the contact medium.

Measures

The advanced level of the Otis-Lennon Mental Ability Test (Otis & Lennon, 1967) was used as a pre-post performance measure. This instrument has been found to exhibit strong test-retest reliability ($r = .94$) over a 1-year period (Smith, 1970). Split-half, Kuder-Richardson, and alternate-forms reliability coefficients also have been found to range from .92 to .96 (Otis & Lennon, 1967). The Otis-Lennon was administered using directions modified to increase subject anxiety. More specifically, these instructions stated that subjects would be rank ordered in terms of the test results and would be provided with their individual standing in comparison to other students. Following the Otis-Lennon, the State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1970) was administered. At posttesting these same two measures were readministered along with the TAS (Sarason, 1957).

Procedure

Subjects in the two treatment and attention-placebo conditions each received six 45-minute individual sessions over a period of 2 weeks. Pretesting and post-testing occurred within 3 days of initiating and terminating treatment, respectively. All subjects for these three conditions were seated in a reclining chair and received relaxation instructions from a cassette tape recording. In preparing the tapes, the narrator was unaware of the hypotheses, measures, or rationale for the study.

EMG-assisted cue-controlled relaxation group (CCR-B). After the subject was seated by the experimenter, electrodes were attached over the frontalis muscle in accordance with the operating manual. The sensitivity-gain scale was set at 25 μ V, and the subject was asked to relax comfortably with eyes closed for approximately 30 sec to establish a baseline of audible feedback. Periodic observations were made during the session through the one-way mirror to insure that continuous feedback was being generated. Once feedback was established, the experimenter began the tape recording and left the room. The tape included three components: (a) a brief statement of rationale for EMG biofeedback training, (b) instructions for progressive relaxation training prepared from Bernstein and Borkovec (1973), and (c) instructions for cue-controlled relaxation

training as presented by Russell and Sipich (1973). More specifically, instructions for the third component directed the subject to focus on his or her breathing and to say the word *relax* with each exhalation. This was continued for approximately 20 pairings. After a 60-sec interval during which the subject was instructed to focus his or her attention on general feelings of relaxation, 20 additional pairings were rehearsed prior to terminating the session. For both cue-controlled relaxation groups, the subject was instructed to use this relaxation technique when faced with daily anxiety-evoking situations, including academic tests.

Cue-controlled relaxation group (CCR). Subjects in this condition were treated in the same manner as subjects in the preceding condition, with the exception that no biofeedback apparatus was used. The tape for this condition was identical to that for the biofeedback group, except that the first component presented a brief rationale for the use of relaxation training only, and all references to biofeedback were deleted.

Attention-placebo group (AP). Subjects in this condition were treated in the same manner as subjects in the CCR group. The tape for this condition, however, consisted primarily of soothing music performed by Roger Williams. A brief rationale for the use of music for relaxation purposes was presented, and at four points during the music the narrator used suggested imagery unrelated to test taking. There were no references during this tape to either biofeedback or cue-controlled relaxation. The AP tape was approximately 37 minutes in length, as compared to 42 minutes for the two other treatment tapes. All three tapes included statements encouraging the subjects to practice their respective treatments daily at home and to use their respective strategies in anxiety-evoking situations. After the initial session, the rationale component of each of the three tapes was deleted for the remaining five sessions.

No-treatment control group (NT). Subjects in this group were told that due to facility limitations, treatment could not be provided at present and that they had been placed on a waiting list for treatment in the future.

Statistical Analysis

Statistical analysis was conducted using gain scores (Huck & McLean, 1975) on the TAS, the STAI State scale (A-State), the STAI Trait scale (A-Trait) and the Otis-Lennon Mental Ability Test. For these dependent variables, two a priori orthogonal comparisons were conducted using t tests. The comparisons included the two treatment conditions combined versus the two control conditions combined and the CCR-B condition versus the CCR condition.

To test for main effects, the dependent measure gain scores were subjected to a series of univariate analyses of variance. Post hoc comparisons, using Scheffé's multiple range test, were conducted for those dependent variables yielding a significant F ratio. The level of significance was set at .05 in all cases.

Table 1
Pre and Post Group Means and Standard Deviations for the Four Dependent Variables

Variable	Treatment condition							
	CCR-B		CCR		AP		NT	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
TAS								
Pre	28.2	3.05	26.7	4.27	26.3	3.12	26.4	2.99
Post	24.2	6.88	18.8	7.98	22.2	3.36	26.1	3.14
A-State								
Pre	49.4	8.30	47.5	14.88	50.8	6.37	48.3	12.15
Post	30.6	7.12	26.6	6.31	34.7	8.08	46.4	10.38
A-Trait								
Pre	48.2	7.13	43.9	12.39	47.3	7.51	44.6	5.30
Post	42.9	6.94	36.2	7.66	39.7	5.54	40.8	6.11
Otis-Lennon								
Pre	46.9	19.70	46.8	12.39	48.6	16.22	47.1	9.0
Post	54.9	15.83	54.9	12.32	51.0	16.39	49.5	9.58

Note. CCR-B = electromyographic-assisted cue-controlled relaxation; CCR = cue-controlled relaxation; AP = attention placebo; NT = no treatment; TAS = Test Anxiety Scale; A-State = State scale of the State-Trait Anxiety Inventory; A-Trait = Trait scale of the State-Trait Anxiety Inventory; Otis-Lennon = Otis-Lennon Mental Ability Test.

Results

Analysis of variance of pretest scores yielded nonsignificant *F* ratios across all dependent variables. Pretest and posttest group means and standard deviations for the dependent variables are presented in Table 1.

Self-report measures of test anxiety and state anxiety indicated that the combined experimental conditions were superior to the combined placebo and no-treatment control conditions. No differences were noted for the self-report measure of trait anxiety. Results

from the Otis-Lennon indicated that the combined treatment conditions made gains on this performance measure that were significantly greater than the combined control conditions. There were no differences, however, between EMG-assisted and traditional cue-controlled relaxation training in terms of any of the dependent measures. The *t* ratios for these *a priori* comparisons are presented in Table 2.

The series of univariate analyses of variance of change scores yielded significant *F* ratios for three dependent variables as follows: TAS, $F(3, 36) = 4.388$, $p < .01$; A-State,

Table 2
A Priori Orthogonal Comparison t Values for the Four Dependent Variables

Comparison	<i>df</i>	TAS	A-State	A-Trait	Otis-Lennon
CCR-B and CCR combined vs.					
AP and NT combined	36	-2.532*	-3.297**	-.460	4.640***
CCR-B vs. CCR	36	1.862	.439	.976	-.581

Note. TAS = Test Anxiety Scale; A-State = State scale of the State-Trait Anxiety Inventory; A-Trait = Trait scale of the State-Trait Anxiety Inventory; Otis-Lennon = Otis-Lennon Mental Ability Test; CCR-B = electromyographic-assisted cue-controlled relaxation; CCR = cue-controlled relaxation; AP = attention placebo; NT = no treatment.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 3

Means, Standard Deviations, and Post Hoc Comparisons for the Change Scores of the Four Dependent Variables

Variable	Treatment condition							
	CCR-B		CCR		AP		NT	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
TAS	-4.0 _{a,b}	5.59	-7.9 _a	5.85	-4.1 _b	4.14	-.3 _b	2.21
A-State	-18.8 _a	12.76	-20.9 _a	14.21	-16.1 _a	8.60	-1.9 _b	4.33
A-Trait	-5.3 _a	4.47	-7.7 _a	6.71	-7.6 _a	6.27	-3.8 _a	4.04
Otis-Lennon	8.0 _a	5.14	8.1 _a	3.24	2.4 _b	3.71	2.4 _b	2.91

Note. Scheffé's multiple comparison test is at the .05 level. Means with the same subscript are not significantly different. CCR-B = electromyographic-assisted cue-controlled relaxation; CCR = cue-controlled relaxation; AP = attention placebo; NT = no treatment; TAS = Test Anxiety Scale; A-State = State scale of the State-Trait Anxiety Inventory; A-Trait = Trait scale of the State-Trait Anxiety Inventory; Otis-Lennon = Otis-Lennon Mental Ability Test.

$F(3, 36) = 6.429$, $p < .01$; and Otis-Lennon, $F(3, 36) = 7.178$, $p < .001$. Post hoc comparisons using Scheffé's test for multiple comparisons indicated that the CCR condition was superior to the no-treatment condition but was no different from the CCR-B or AP conditions for the TAS. In terms of state anxiety the CCR, CCR-B, and AP conditions were found superior to the no-treatment condition. No differences were found between conditions for trait anxiety. Both the CCR and CCR-B groups were found to result in significantly greater gains on the Otis-Lennon when compared to either control condition, but they were not significantly different from each other. Without exception, subjects in the experimental conditions increased performance at posttesting. These increases ranged from 2 to 18 points. For the two control conditions, however, these gains were found to range from 1 to 8 points only. Furthermore, 30% of the subjects in the control conditions actually demonstrated a decrease in performance at posttesting. Means and standard deviations for the change scores for the four dependent variables by treatment condition are presented in Table 3.

Discussion

As hypothesized, the combined cue-controlled relaxation conditions were found to reduce self-reported test anxiety and state

anxiety more than the placebo and no-treatment conditions. Also, the combined experimental conditions resulted in a greater increase in test-taking performance than the combined control conditions. However, the hypothesis that EMG-assisted cue-controlled relaxation would be superior to traditional cue-controlled relaxation was not supported.

The findings indicate that these approaches to cue-controlled relaxation may be effective for the treatment of test anxiety. However, despite the considerable enthusiasm generated around the use of EMG biofeedback in relaxation training, it appears to add little to the effectiveness of cue-controlled relaxation. The general assumption that frontalis relaxation via EMG training generalizes to other muscle groups may not be totally accurate. Alexander (1975) found little evidence to support the use of EMG biofeedback alone in inducing generalized muscular relaxation. Given the potential cost and inconvenience of using EMG training, it may not be the treatment of choice for approaches using generalized relaxation training. The area in which biofeedback is making its most important contribution appears to be in the field of specific psychophysiological disorders (Budzynski, Stoyva, & Adler, 1970; Budzynski, Stoyva, Adler, & Mullaney, 1973).

Of particular interest in this study is the significant increase in test-taking performance for the subjects in the experimental conditions. The mean gains on the Otis-Lennon for these

subjects translated into an increase of between 7 and 24 IQ points, depending on the location of the scores in the distribution. The mean gain scores for the combined control conditions translated into an increase of between 1 to 4 IQ points only. The literature dealing with the modification of test anxiety is replete with research yielding significant differences in self-reported test anxiety. Increases on performance measures, however, are not as plentiful. Finger and Galassi (1977) stated that a review of the literature revealed that performance improvements were obtained in only 16 (29.6%) of 54 studies. Although the present study resulted in significant increases in scores on a test of mental ability, replication and further investigation using performance measures is needed.

It is noteworthy that we found no differences in the reduction of anxiety between the experimental treatments and the attention-placebo condition on any self-report measure. This would emphasize further the need for performance as well as self-report measures in the assessment of test anxiety.

There are several important questions that remain unanswered. For one, what are the differential treatment effects for individual subjects receiving EMG-assisted cue-controlled relaxation? It may be that subjects who are trained successfully to modify EMG frontalis responses will demonstrate an increase in performance, whereas those who fail to do so will not. Research designed to answer this question is needed.

Implications for further research with cue-controlled relaxation itself include investigating the relative contributions of various approaches to relaxation training and various forms of the subvocalized cue words. The possible inclusion of additional subvocalized statements as a supplement to the cue word appears promising. In this manner, many of the task-orienting self-verbalizations recommended by Meichenbaum (1977) could be incorporated into the cue-controlled relaxation treatment. Cognitive behavior modification approaches, without relaxation, could also be compared to cue-controlled relaxation. In this manner the relative importance of relaxation training in the treatment of test anxiety could be assessed.

References

- Alexander, A. B. An experimental test of assumptions relating to the use of electromyographic biofeedback as a general relaxation technique. *Psychophysiology*, 1975, 12, 656-662.
- Alexander, A. B., French, C. A., & Goodman, N. J. A comparison of auditory and visual feedback in biofeedback assisted muscular relaxation training. *Psychophysiology*, 1975, 12, 119-123.
- Bernstein, D. A., & Borkovec, T. D. *Progressive relaxation training: A manual for the helping professions*. New York: Research Press, 1973.
- Budzynski, T. H., Stoyva, J. M., & Adler, C. S. Feedback-induced muscle relaxation: Application to tension headache. *Journal of Behavior Therapy and Experimental Psychiatry*, 1970, 1, 205-211.
- Budzynski, T. H., Stoyva, J. M., Adler, C. S., & Mullaney, D. J. EMG biofeedback and tension headache: A controlled outcome study. *Psychosomatic Medicine*, 1973, 35, 484-496.
- Canter, A., Kondo, C. Y., & Knott, J. R. A comparison of EMG feedback and progressive muscle relaxation training in anxiety neurosis. *British Journal of Psychiatry*, 1975, 127, 470-477.
- Chang-Liang, R., & Denney, D. R. Applied relaxation as training in self-control. *Journal of Counseling Psychology*, 1976, 23, 183-189.
- Finger, R., & Galassi, J. P. Effects of modifying cognitive versus emotionality responses in the treatment of test anxiety. *Journal of Consulting and Clinical Psychology*, 1977, 45, 280-287.
- Goldfried, M. R., & Trier, C. S. Effectiveness of relaxation as an active coping skill. *Journal of Abnormal Psychology*, 1974, 83, 348-355.
- Haynes, S. N., Moseley, D., & McGowan, W. T. Relaxation training and biofeedback in the reduction of frontalis muscle tension. *Psychophysiology*, 1975, 12, 547-552.
- Huck, S. W., & McLean, R. A. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, 1975, 82, 511-518.
- Johnson, S. M., & Sechrest, L. Comparison of desensitization and progressive relaxation in treating test anxiety. *Journal of Consulting and Clinical Psychology*, 1968, 32, 280-286.
- Marchetti, A., McGlynn, F. D., & Patterson, A. S. Effects of cue-controlled relaxation, a placebo treatment, and no treatment or changes in self-reported and psychophysiological indices of test anxiety among college students. *Behavior Modification*, 1977, 1, 47-72.
- Meichenbaum, D. *Cognitive behavior modification: An integrative approach*. New York: Plenum Press, 1977.
- Otis, A. S., & Lennon, R. T. *Otis-Lennon Mental Ability Test: Manual for administration*. New York: Harcourt, Brace & World, 1967.
- Reeves, J. L., & Mealiea, W. L. Biofeedback-assisted cue-controlled relaxation for the treatment of flight phobias. *Journal of Behavior Therapy and Experimental Psychiatry*, 1975, 6, 105-109.
- Reinking, R. H., & Kohl, M. L. Effects of various forms of relaxation training on physiological and self-

- report measures of relaxation. *Journal of Consulting and Clinical Psychology*, 1975, 43, 595-600.
- Russell, R. K., Miller, D. E., & June, L. N. Group cue-controlled relaxation in the treatment of test anxiety. *Behavior Therapy*, 1974, 5, 572-573.
- Russell, R. K., Miller, D. E., & June, L. N. A comparison between group systematic desensitization and cue-controlled relaxation in the treatment of test anxiety. *Behavior Therapy*, 1975, 6, 172-177.
- Russell, R. K., & Sipich, J. F. Cue-controlled relaxation in the treatment of test anxiety. *Journal of Behavior Therapy and Experimental Psychiatry*, 1973, 4, 47-49.
- Russell, R. K., & Sipich, J. F. Treatment of test anxiety by cue-controlled relaxation. *Behavior Therapy*, 1974, 5, 673-676.
- Russell, R. K., Wise, F., & Stratoudakis, J. P. Treatment of test anxiety by cue-controlled relaxation and systematic desensitization. *Journal of Counseling Psychology*, 1976, 23, 563-566.
- Sarason, I. G. Test anxiety, general anxiety and intellectual performance. *Journal of Consulting Psychology*, 1957, 21, 485-490.
- Smith, A. E. Test reviews (Review of Otis Lennon Mental Ability Test by A. S. Otis & R. T. Lennon). *Journal of Counseling Psychology*, 1970, 17, 91-92.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *Manual for the State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Wickramasekera, I. Instructions and EMG feedback in systematic desensitization: A case report. *Behavior Therapy*, 1972, 3, 460-465.

Received July 21, 1977 ■

Vicarious Anxiety: A Systematic Evaluation of a Vicarious Threat to Self-Esteem

Philip C. Kendall
University of Minnesota

A. J. Finch, Jr.
Virginia Treatment Center for Children
Richmond, Virginia

L. E. Montgomery
Middle Peninsula-Northern Neck Mental Health and Mental Retardation Services
Saluda, Virginia

This research investigated, in three studies, subjects' state anxiety arousal in response to an in vivo vicarious threat to self-esteem. In Studies 1 and 2 students observed a guest speaker who provided the anxiety manipulation. In both studies, correlation and median split analyses indicated that high empathic and low trait anxious subjects reported elevated state anxiety in response to the vicarious threat. Even though external subjects reported overall higher levels of state anxiety, no differential responsiveness between internal and external subjects was found. When subjects were *matched* on initial state anxiety, high-empathy subjects were found to have experienced vicarious anxiety, whereas subjects low on empathy did not. An analysis of high and low trait anxious subjects who were matched on initial state anxiety did not reveal differential responsiveness. In addition to replicating Study 1, Study 2 found that the *Helplessness: an inability to affect change in others* factor of locus of control was significantly negatively related to empathy, and the cognitive reappraisal styles of reversal (denial, reaction formation) and projection were related to state anxiety *decreases*. Study 3 provided evidence for the absence of a confound.

Imagine an evening "on the town" in which an audience is exposed to a theatrical misfortune—an unprepared understudy. As anyone who has been in such a situation knows, there will be large differences in the behavior of the ticket holders. Some will demand a refund, whereas others will feel sorry and work themselves into an anxious sweat! What are the relevant personality characteristics and cognitive reappraisal styles that are associated with such response variation?

Present psychological evidence tends to support the notion that virtually all learning

phenomena that are directly experienced can also occur vicariously through the observation of another person (Bandura, 1969). The majority of the evidence demonstrates that such observation affords the acquisition of response patterns and the therapeutic benefits of a coping model. In the present study observational or vicarious emotionality (anxiety), a less desirable response pattern, was investigated.

Those investigations that have in the past targeted emotionality have usually been concerned with such variables as the type of stress (physical danger threat or threat to self-esteem: Bennett & Holmes, 1975; Kendall, 1978; Kendall, Finch, Auerbach, Hooke, & Mikulka, 1976; Speisman, Lazarus, Mordkoff, & Davison, 1964) and the nature of the stress experience (direct or vicarious: Alfert, 1966, 1967; Averill, Olbrich, & Lazarus, 1972; Opton & Lazarus, 1967), whereas others have in-

Portions of the present article were prepared while the first author was at the Center for Advanced Study in the Behavioral Sciences, Stanford, California. This research was initiated while the first author was affiliated with Virginia Commonwealth University.

Requests for reprints should be sent to Philip C. Kendall, Department of Psychology, Elliott Hall N422, 75 East River Road, University of Minnesota, Minneapolis, Minnesota 55455.

cluded an examination of threat reduction via defensive styles (Bennett & Holmes, 1975; Holmes & Houston, 1974; Lazarus, Opton, Nomikos, & Rankin, 1965; Speisman et al., 1964). Uniformly, these studies have been conducted in the laboratory setting.

When one organizes the studies in this area and examines a matrix of stress type (physical, ego threat) and nature of stress (vicarious, direct) variables, there is an absence of an evaluation of observational ego threat or vicarious stress to self-esteem. The intent of the present investigation was to conduct such a study and to examine the characteristics of subjects who would experience vicarious anxiety. Three personality measures were selected: empathy (Hogan, 1969), locus of control (Nowicki & Duke, 1974), and anxiety (Spielberger, Gorsuch, & Lushene, 1970).

It was hypothesized that subjects scoring high on empathy, indicating a high degree of sensitivity to the feelings and needs of others, would respond with increases in their personal state anxiety greater than those of the low-empathic subjects following a vicarious threat to self-esteem. Regarding locus of control (i.e., internal control reflecting a generalized expectancy that outcomes or reinforcement are a consequence of one's own behavior, and external control suggesting an expectancy that outcomes are the result of luck, fate, chance, or powerful others), the empirical findings (see Phares, 1973) support the notion that externals are more anxious than internals (Ray & Katahn, 1968; Watson, 1967), and, correspondingly, it was hypothesized that external subjects would be more likely to respond with vicariously aroused state anxiety than internal subjects. Finally, in line with the state-trait theory of anxiety (Spielberger, 1972), which posits that high trait anxious subjects are more prone to experience elevated states of anxiety, it was predicted that high trait anxious subjects would respond with greater state anxiety than low trait anxious individuals to the vicarious threat to self-esteem.

To reduce the unwanted variability due to expectancy and role playing, an *in vivo* situation was selected. Thus, the present study set out to directly and intentionally evaluate subjects reporting state anxiety reactions in a

naturally occurring stimulus condition that includes a vicarious threat to self-esteem.

Study 1

Method

Subjects

The subjects in this study were 30 undergraduate psychology students enrolled in an evening introductory class at an urban Virginia university. There were 14 males and 16 females, with a mean age of 24.5.

Measures

Empathy. Hogan's (1969) empathy scale is a 64-item inventory that requires subjects to endorse true or false items as they apply and measures an individual's sensitivity to the needs of others.

Locus of control. The Adult Nowicki-Strickland Internal-External Scale (ANSIE; Nowicki & Duke, 1974) is a 40-item true-false inventory that assesses an individual's belief in personal control over reinforcements (internal) or that luck, fate, chance, or powerful others are the source of reward (external).

Anxiety. The State-Trait Anxiety Inventory (STAI; Spielberger et al., 1970) consists of two separate 20-item self-report scales for measuring state anxiety (A-State) and trait anxiety (A-Trait). The STAI A-State scale requires people to describe how they feel at a particular moment in time; the STAI A-Trait scale asks people to describe how they generally feel. The A-State scale assesses anxiety at the moment in a given situation, whereas the A-Trait scale measures an individual's global predisposition to feel anxious.

Procedure

Subjects were distributed the empathy and locus of control scales and were required to complete and return them to their instructor. The STAI was administered according to standard instructions during a lecture period under nonstress conditions. Subjects were merely informed that the instructor was "collecting some research data and needed for everyone to fill them out." They were also informed that their scores would not be individually analyzed but that the information would be grouped together for comparisons.

During a later lecture period, the subjects were informed that there would be a guest speaker¹ who would talk on "child psychopathology" and that the speaker was a recognized professional who had published numerous articles in the area. In fact, the guest speaker was a confederate who would perform the experimental manipulation. The manipulation consisted of an introduction by the instructor and a general

¹ The guest speaker was in fact a recognized clinical child psychologist who participated as a confederate and who had prerehearsed the anxiety manipulation.

statement by the guest speaker stating the topic of his talk and that the class would be asked to fill out a questionnaire used in related research as a demonstration. Then a series of events followed, including brief and intermittent stuttering, repeating phrases, mixing ideas, spilling coffee, dropping papers, misplacing slides, and an undelivered slide projector. The series of events lasted between 6 and 7 minutes. When the slide projector could not be located, the confederate guest speaker suggested that the class fill out the demonstration questionnaire while he and the instructor would try to get the slides and projector organized. The STAI was distributed, and subjects were instructed to read the instructions and complete the inventory.

Debriefing

When all subjects had completed the STAI, the manipulation was explained. The guest speaker then presented his talk normally while the instructor and an assistant examined the data. At the end of the guest talk, a brief summary of the results of the present experiment was discussed with the class.²

Results

The change in subjects' A-State from the initial nonstress phase to the postmanipulation phase was examined in a series of analyses for each of the personality measures. In each case, the correlations of the personality measure with the initial nonstress A-State score and the A-State increase score (postmanipulation A-State minus initial A-State) are presented. The correlations are followed by separate 2 (personality variable: median split) \times 2 (trials: premanipulation and postmanipulation) analyses of variance. In addition, when sufficient subjects could be included, subjects *matched* on their initial A-State scores were subjected to 2 \times 2 analyses of variance³ similar to those conducted on all subjects using the median split.

Empathy

The correlation between empathy scores and the initial A-State scores was $-.26$, which only approached significance. However, empathy was significantly correlated with an increase in A-State scores ($r = .39$, $p < .05$). This significant correlation indicates that empathy is directly related to reported elevations in A-State.

When subjects were divided at the median

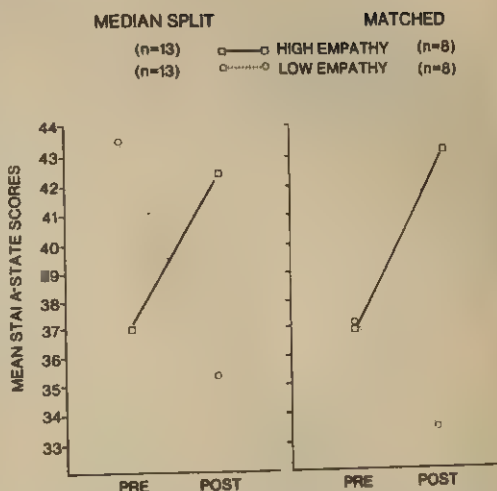


Figure 1. Mean State-Trait Anxiety Inventory State Anxiety (STAI A-State) scores for high- and low-empathy subjects at premanipulation and postmanipulation. (The left portion includes all subjects divided at the median; the right portion includes certain subjects matched on initial A-State scores.)

(38) into high-empathy (>38) and low-empathy (<38) groups (both $n_s = 13$; 4 subjects whose scores equaled 38 were eliminated), the analysis of variance indicated nonsignificant main effects for empathy level and trials, $F_s(1, 24) < 1$. The Trials \times Empathy interaction was significant, $F(1, 24) = 14.27$, $p < .005$, indicating that the A-State scores of high-empathic subjects increased due to the manipulation, whereas the A-State scores of the low-empathic subjects decreased (see lefthand portion of Figure 1). The means of the A-State scores for subjects high and low on empathy were pre = 37.0 and 43.69 and post = 42.46 and 35.15, respectively. As shown in Figure 1, high-empathy subjects were at higher A-State levels following the manipulation than they were at the premanipulation phase. In contrast, low-empathy

² Although several students reported feeling very badly for the guest speaker (one cleaned up the coffee during the talk), the class unanimously stated that they were unaware that it was a manipulation. Importantly, the class also volunteered that the debriefing which included an explanation of some of the results of the study was a very stimulating and innovative teaching method.

³ Subjects' scores were considered matched within a 1-point range. A matched group analysis of variance was not used.

subjects were less anxious following the manipulation. Thus, subjects who scored high on empathy were those who reported a vicarious anxiety response.

However, a reevaluation of the left-hand portion of Figure 1 reveals that the initial A-State scores for the two groups differ substantially. (Recall also the $-.26$ correlation between empathy and initial A-State.) Anxiety is not entirely independent of empathy, and legitimate inferences about the effect of empathy on state anxiety can be drawn only when empathy is varied independently of anxiety.⁴ To achieve independence, 16 subjects from the high- and low-empathy groups were matched according to their initial A-State scores, and changes in anxiety were analyzed. The premanipulation and postmanipulation means for the matched high- and low-empathy groups were pre = 36.75 and 37.25 and post = 43.25 and 33.37, respectively. Results of a 2 (empathy level) \times 2 (trials) analysis of variance indicated nonsignificant main effects for empathy level and for trials, $F_s(1, 14) < 1$, but a significant interaction, $F(1, 14) = 10.91$, $p < .001$. This finding is presented in the right-hand portion of Figure 1. Here, the high-empathy subjects reported increased levels of state anxiety following the manipulation, and the low-empathy subjects reported decreased anxiety. The results of the matched-subjects analysis coincide with those reported for the median-split procedure and demonstrate more clearly the effects of empathy on A-State changes.

Locus of Control

The correlation between locus of control and the initial A-State score was $.39(p < .05)$, and the correlation between locus of control and the increase in A-State scores was $.09(p > .10)$. These findings support those of Ray and Katahn (1968) and Watson (1967), who reported that external subjects score higher on anxiety than internals. However, the present evidence does not support the hypothesis that externals would be more likely to experience vicarious anxiety.

When subjects were divided at the median (9) into internal ($n = 11$, scores < 9) and external ($n = 9$, scores > 9) locus of control

groups (10 subjects whose scores were 9 were eliminated), the results of a 2 (internal/external) \times 2 (trials) analysis of variance indicated a significant main effect of locus of control, $F(1, 18) = 11.61$, $p < .005$, a nonsignificant main effect for trials, $F(1, 18) = 1.04$, and a nonsignificant interaction, $F(1, 18) = 1.59$. Examination of the means for the internals and externals (pre = 35.09 and 44.22 and post = 29.27 and 46.22, respectively) indicated that the significant internal/external difference in A-State was reflected in the higher A-State scores for externals. An attempt to analyze for the effects of locus of control on state anxiety with subjects matched on initial A-State was aborted by the limited number (i.e., three) of matched cases.

Anxiety

The correlation between A-Trait and the initial nonstress A-State was $.59(p < .01)$, and the correlation between increases in A-State and A-Trait scores was $-.48(p < .01)$. The significant A-Trait/A-State correlation is not surprising; however, the significant negative correlation of A-State increases and A-Trait scores was not expected. This unexpected finding suggests that it was the lower A-Trait scores that were related to the greater state anxiety increases.

When subjects were divided at the median of their A-Trait scores (38), high A-Trait subjects ($n = 12$) had scores greater than 38, whereas low A-Trait subjects ($n = 12$) had scores less than 38 (6 subjects were eliminated by the median split). The mean A-State scores for high and low A-Trait groups were pre = 45.83 and 31.50 and post = 38.83 and 34.25, respectively. Results of a 2 (A-Trait level) \times 2 (trials) analysis of variance indicated a significant A-Trait level effect, $F(1, 22) = 13.38$, $p < .005$, and a nonsignificant trials effect, $F(1, 22) = 1.04$. The A-Trait Level \times Trials interaction was significant, $F(1, 22) = 5.49$, $p < .05$, indicating differential changes in A-State for subjects differing in A-Trait (see Figure 2). Figure 2 indicates that the low A-Trait subjects were more state anxious

⁴ The authors wish to thank an anonymous reviewer for stressing this important issue.

following the vicarious threat to self-esteem than prior to it, whereas the high A-Trait subjects reported less A-State following the manipulation. It should be noted here that the extreme scores at the premanipulation phase set limits on the probable size and direction of any change, and an analysis of subjects *matched* on initial A-State would be more desirable. Given that A-Trait and A-State are not independent, it was not surprising that only 3 subjects could be matched, and thus no matched-subjects analysis was conducted.

Discussion

The results of the present study provide strong support for the hypothesis that high-empathic subjects will respond anxiously after observing an anxious speaker. Correspondingly, the present study provides validation for Hogan's (1969) empathy scale. In a discriminant fashion, the locus of control results are also valuable. That is, locus of control appears to be unrelated to vicarious emotionality. However, external subjects did report more anxiety, as has been found in other studies.

The results of the trait anxiety analysis were not entirely as expected. Although the significant interaction of high/low trait level and trials was predicted, the finding that the low A-Trait subjects increased while high A-Trait subjects decreased in state anxiety as a function of the manipulation was *contradictory* to our hypothesis. Indeed, this is the first known study to disclose a situation in which the low A-Trait individuals were more disposed to be state anxious, and the findings are directly contradictory to those of numerous studies using different stresses (e.g., Auerbach, 1973; Hodges & Spielberger, 1966) and to the definition of A-Trait itself. A finding as surprising as this appeared to warrant replication.

Yet another potential need for replication concerns the multidimensionality of locus of control. Recent evidence using the Rotter scale (Abramowitz, 1973; Levenson, 1973) and the Nowicki-Strickland children's form (Kendall, Finch, & Mahoney, 1976) have shown that although overall locus of control is predictive, factor-specific scoring of the dimen-

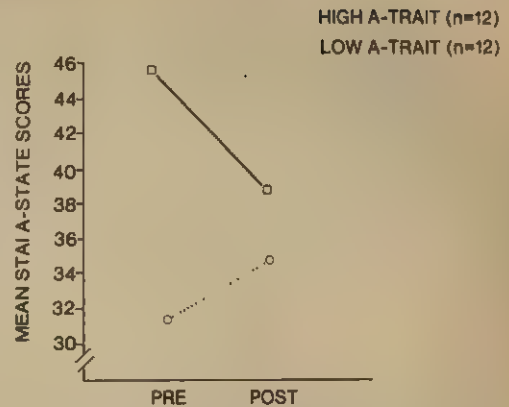


Figure 2. Mean State-Trait Anxiety Inventory State Anxiety (STAI A-State) scores for all high and low Trait Anxiety (A-Trait) subjects at premanipulation and postmanipulation.

sions of locus of control can advance accurate hypotheses. The factor analytic demonstration of the multidimensionality of the ANSIE (Kendall, Finch, & Mikulka, Note 1) produced five meaningful factors. Among the factors were two representing helplessness—one that appeared to be related to an absence of alternatives or solutions and another that is suggestive of an inability to affect change in others. Another factor expressed superstitious beliefs. It was felt that perhaps, even though overall locus of control was not related to vicarious anxiety, the specific factors might be related. Specifically, the *Helplessness: An inability to affect change in others* factor of the ANSIE was hypothesized to be negatively related to the vicarious anxiety experience. In the present study, subjects watched an individual anxiously blunder through a professional presentation. Those who were sensitive to the needs and feelings of others (i.e., high on empathy) showed anxiety arousal. Similarly, it might be hypothesized that individuals who perceive themselves as being capable of affecting change in others might also be aroused.

Finally, an examination of the means of the high- and low-empathy and high and low A-Trait subjects revealed that although high-empathic and low A-Trait subjects reported increased state anxiety, the mean scores of both the low-empathic and high A-Trait subjects *decreased* after the manipulation. This

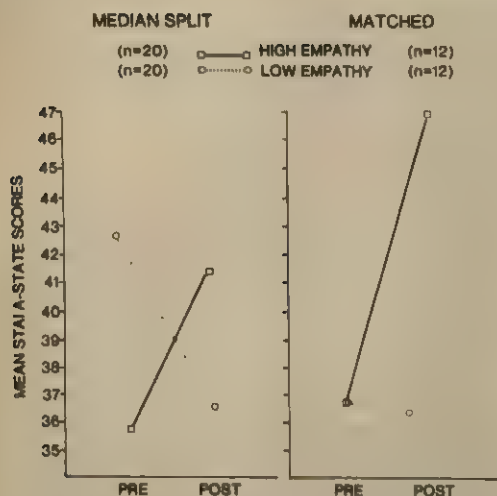


Figure 3. Mean State-Trait Anxiety Inventory State Anxiety (STAI A-State) scores for high- and low-empathy subjects at premanipulation and postmanipulation. (The left portion contains all subjects divided at the median; the right portion includes certain subjects matched on initial A-State scores.)

decrease suggests that perhaps certain cognitive reappraisals (Lazarus, 1966) had been used by these subjects and that these cognitive styles (i.e., defense mechanisms) should be examined. For this, and the above reasons, a replication and extension was conducted.

Study 2

Method

The subjects in this study were 40 undergraduate psychology students enrolled in an early morning introductory psychology class at an urban Virginia university. There were 20 males and 20 females with a mean age of 19.7.

Measures

In addition to the measures used to replicate the initial study, the Defense Mechanisms Inventory (DMI, Gleser & Ihilevich, 1969) was administered. The DMI assesses the relative intensity of five major groups of defenses. The inventory consists of 10 brief stories followed by four questions regarding the individual's actual behavior, fantasy behavior, thoughts, and feelings in the situation described. The five defense groups are (a) hostility-out (i.e., displacement), (b) projection, (c) principalization (i.e., intellectualization, rationalization, isolation), (d) turning-against-self, and (e) reversal (i.e., negation, denial, reaction formation). Validity data are provided in Gleser and Ihilevich (1969) and Gleser and Sacks (1973).

Procedure

The same experimental procedure was followed as in Study 1. Similarly, subjects were debriefed and were presented with a brief outline of the results.

Results and Discussion

As in Study 1, the A-State reaction of subjects to the vicarious anxiety manipulation was examined in a series of analyses (i.e., correlational, median-split analysis of variance, and analysis of subjects matched on initial A-State) for each of the personality measures.

Empathy

The correlation between empathy and the initial nonstress measure of A-State was $-.28$, which approached but did not reach significance. However, empathy was significantly related to A-State increases ($r = .35$). As in Study 1, the more empathic subjects reported elevated state anxiety in response to the vicarious anxiety manipulation.

When subjects were again split at the median (38) into high-empathy (>38 , $n = 20$) and low-empathy (<38 , $n = 20$) groups, the results of the analysis of variance indicated nonsignificant main effects for high/low empathy and trials, $F(1, 38) < 1$, and $F(1, 38) < 1$, respectively. The interaction of empathy and trials was significant, $F(1, 38) = 5.29$, $p < .03$, indicating that the state anxiety level of subjects high on empathy changed differentially from that of subjects low on empathy. Again, as in Study 1, high-empathic subjects showed an increase in A-State due to the manipulation, whereas the subjects low on empathy did not. This interaction is presented in the left-hand portion of Figure 3. In addition, the extent of the replication of Study 1 should be noted. (See left-hand portion of Figure 1.)

To examine the effects of empathy on state anxiety when empathy is varied independently of anxiety, 24 subjects from the high- and low-empathy groups were matched according to their initial A-State scores. The premanipulation and postmanipulation means for the matched high- and low-empathy groups were, respectively, pre = 36.72 and 36.72 and post = 46.1 and 36.40. Results indicated nonsignificant main effects, $F(1, 22) = 1.69$, and

$F(1, 22) = 1.48$, respectively, and a significant Empathy Level \times Trials interaction, $F(1, 22) = 4.37$, $p < .05$. This finding is presented in the right-hand portion of Figure 3. As can be seen, when high- and low-empathy subjects are matched on initial A-State, it is the high-empathic individuals who report experiencing vicariously aroused anxiety. It should be noted that this finding is quite similar to that found in the matched-subjects analysis in Study 1 (compare to right-hand portion of Figure 1), with the exception that the low-empathic subjects in Study 2 did not show a marked reduction in anxiety. Nevertheless, the results regarding empathy in Figures 1 and 3 represent quite similar findings. Indeed, the results of the present study provide strong, replicated support for the finding that high-empathic individuals experience vicarious anxiety arousal, whereas less empathic people do not. Moreover, the aroused anxiety due to a vicarious threat to self-esteem in the high-empathic subjects also provided validation for the empathy scale (Hogan, 1969). The state anxiety increases reported by high-empathic subjects are considered the result of their placing themselves in the position of the speaker—"If I were up there, I would be extremely anxious" or "I can imagine what he feels like." It would be interesting to examine in future research the self-statements made by empathic and nonempathic observers and their ensuing emotional response patterns.

Locus of Control

Two types of analyses were conducted using the locus of control data: analyses of overall locus of control scores and analyses of the separate factors within the locus of control scale.

Overall score analyses. The correlation between locus of control and the initial nonstress measure of A-State was .31 ($p < .05$), whereas the relationship to A-State increases was nonsignificant ($r = -.15$). As in Study 1, externals reported higher A-State, but locus of control remained unrelated to the vicarious arousal of anxiety.

Subjects were divided at the median of the overall scores (9) into internal ($n = 18$, scores < 9) and external ($n = 18$, scores > 9) groups

(4 subjects who scored 9 were eliminated) with subsequent analyses of A-State scores across trials. Results indicated nonsignificant main effects and a nonsignificant interaction, $F(1, 34) = 3.40$, and $F(1, 34) < 1$, $F(1, 34) = 1.47$, respectively. The means for the internal and external subjects, respectively, were pre = 35.16 and 43.16 and post = 38.33 and 39.27. The locus of control main effect approached significance ($p < .07$) in the direction similar to Study 1 in which externals reported more state anxiety.

To examine the effects of locus of control on A-State when locus of control is varied independently of anxiety, 22 subjects from the internal and external groups were matched on initial A-State. The premanipulation and postmanipulation means for matched internal and external subjects, respectively, were pre = 35.54 and 35.90 and post = 37.45 and 38.09. The results of the analysis of variance produced nonsignificant main effects for locus of control and trials and a nonsignificant interaction, $F(1, 20) = 1.76$, $F(1, 20) = 1.96$, and $F(1, 20) < 1$, in that order. More conclusively, these findings demonstrate that locus of control is not related to vicarious anxiety arousal.

Factor-specific analyses. Examination of the scores on the specific factors of locus of control revealed no additional meaningful relationships. That is, neither the overall locus of control score nor the five factor-specific scores were correlated with an increase in state anxiety (all r s $< .27$). Thus, the factor scores did not provide additional specificity.

On the other hand, although the overall locus of control score did not correlate significantly with the empathy measure ($r = .19$), Factor II, *Helplessness: an inability to affect change in others*, did correlate meaningfully ($-.35$, $p < .05$). This relationship suggests that subjects who are empathic perceive themselves as having the ability to affect change in others.

Anxiety

Trait anxiety correlated .61 ($p < .005$) with initial A-State and $-.52$ ($p < .005$) with increases in A-State. These results support those of Study 1 and, specifically, support the

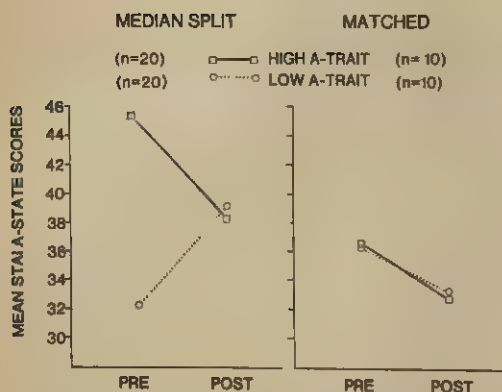


Figure 4. Mean State-Trait Anxiety Inventory State Anxiety (STAI A-State) scores for high and low Trait Anxiety (A-Trait) subjects at premanipulation and postmanipulation. (The left portion contains all subjects divided at the median; the right portion includes certain subjects matched on initial A-State scores.)

surprising finding that it was the lower A-Trait subjects who responded with increases in A-State.

A 2 (A-Trait level) \times 2 (trials) analysis of variance of subjects divided at the median of their A-Trait scores (38) into high ($n = 20$, A-Trait > 38) and low ($n = 20$, A-Trait < 38) A-Trait groups revealed a significant main effect for A-Trait level, $F(1, 38) = 6.01$, $p < .02$, a nonsignificant main effect for trials, $F(1, 38) < 1$, and a significant interaction, $F(1, 38) = 7.75$, $p < .005$. The means for high and low A-Trait subjects were pre = 45.83 and 31.50 and post = 38.83 and 34.25. Again, similar to Study 1, the significant Trait Anxiety Level \times Trials interaction was examined and was found to indicate that the low A-Trait subjects had reported more state anxiety following the manipulation than at a nonstress period, whereas the high A-Trait subjects had reported less. This interaction is presented in the left-hand portion of Figure 4. It should be noted again that the initial A-State means were very different for the A-Trait groups and that these initial score differences set limits on the possible size and direction of change.

A more crucial examination of the effects of trait anxiety on state anxiety was conducted using 20 subjects matched on initial A-State. The premanipulation and postmanipulation means were high A-Trait = 36.9 and 33.2 and

low A-Trait = 36.5 and 33.6. An analysis of variance revealed nonsignificant differences, $F(1, 18) < 1$, $F(1, 18) = 1.29$, and $F(1, 18) < 1$, for A-Trait level, trials, and the interaction, in that order. Thus, the replicated finding in Studies 1 and 2 that high and low A-Trait subjects responded differently to the vicarious threat to self-esteem was not supported by the analysis of matched subjects. (See right-hand portion of Figure 4.)

Future research should consider situation-specific anxiety traits (Endler & Okada, 1975; Zuckerman, 1977) and the prediction of arousal based on the congruence of the class of the anxiety trait measure and that of the experimental situation (Kendall, 1978). Predictions based on the interactional assessment of trait anxiety should be quite heuristic in relation to vicarious anxiety.

Age and Sex

An examination of the correlations between age, sex, and the increase in state anxiety indicated that neither age ($r = -.25$) nor sex ($r = -.14$) were significantly related to A-State increases. In addition, neither sex nor age was related to empathy (both r s $< .08$), nor was sex related to trait anxiety ($r = .06$). Age was related to trait anxiety ($r = .33$, $p < .05$).

Cognitive Reappraisals/Defense Mechanisms

To examine the reappraisal styles of subjects who did not become anxious following our manipulation (some subjects actually de-

Table 1
Correlations of the Defense Mechanism Inventory Categories With Decreases in State Anxiety and With Locus of Control

Category	A-State decrease	Locus of control
Hostility-out	.25	.36*
Projection	.34*	.22
Principalization	.02	-.31*
Turning-against-self	.15	.08
Reversal	.42**	-.40**

* $p < .05$.

** $p < .01$.

creased in reported state anxiety), the five scores on the DMI were intercorrelated with decreases in state anxiety (initial A-State minus postmanipulation A-State). The results are presented in Table 1. As can be seen, there exist significant relationships between subjects who characteristically use reversal and projection and their changes in state anxiety. These relationships suggest that subjects who characteristically use reversal-type defenses (i.e., denial, negation, reaction formation) and projection are likely to report a nonanxious behavior pattern when exposed to an anxious model. Here, an observer using a reversing reappraisal style (e.g., denial) might think, "He's not really anxious" or an observer might project the emotionality onto another person—"It looks like the teacher is getting uncomfortable." Future research should examine the content of the self-statements made during cognitive reappraisals of threat.

Although four of the DMI categories did not correlate significantly with trait anxiety, there was a significant relationship between reversal and trait anxiety ($r = .51, p < .005$), suggesting that the high A-Trait subjects tended to rely on the reversal defensive styles. None of the defensive styles correlated with empathy.

Locus of control and defense mechanisms. Although this is an ancillary section of the present study, several findings are noteworthy. The correlations of interest were included in Table 1. As can be seen, a significant relationship between locus of control and the defense styles of reversal, hostility-out, and principalization was found. These relationships suggest that internal subjects are more characteristic users of reversal and principalization, whereas externals tend to rely on hostility-out (i.e., displacement).

The results of Study 2 suggest that the state anxiety of subjects whose style of cognitive reappraisal emphasizes reversal and projection was reduced by the anxious model. This is consistent with the findings of Houston (1971, 1973), Houston and Hodges (1970), and Lazarus and Alfert (1964), who reported the advantage of denial in stressful situations and the arousal-reducing qualities of denial and reaction formation (reversals). The present findings are supportive of Bennett and Holmes

(1975), who reported a reduction in arousal relating to projection, but are in contrast to Holmes and Houston (1971) and Stevens and Reitz (1970), who did not support the projection/anxiety reduction hypothesis. The reversal defensive mechanisms of the DMI include denial, reaction formation, and negation. The present findings support the relationship of such reversing as a style of cognitive reappraisal to obliterate anxiety arousal.

Comment on the Replication

The results of the first two studies provide replicated evidence that state anxiety as self-reported by high-empathic subjects increases following a vicarious threat to self-esteem manipulation. Trait anxiety was consistently related to both A-State and A-State increases, but A-Trait was not related to A-State increases when subjects were matched on initial A-State. In addition, internal and external subjects do not differ in their tendency to experience vicarious anxiety, but externality is related to higher levels of state anxiety.

The contexts of replicated findings are noteworthy, and the present replication is no exception. In fact, there were actually several differences between Study 1 and Study 2. First, one was an evening class, whereas the other was an 8:00 a.m. class. Second, the mean ages were 24.5 and 19.7 years, respectively. Finally, the first guest speaker for the subjects in Study 1 was the experimental manipulation, whereas, due to other commitments, there were two guest speakers prior to the manipulation for subjects in Study 2. Nonetheless, even with the contextual variations (notwithstanding the care taken to replicate the procedural matter), the results were consistent.

Study 3

Before attempting to draw any conclusions or implications from the results of the first two studies, it must first be asked whether the state anxiety changes were the result of our manipulation. According to Sarason (1972),

models, teachers, and experimenters provide two types of information: 1) what they do or say, and 2) how they do or say it. Both require intensive inquiry because neglect of one of these dimensions could nullify effects of the other. (p. 399)

Even though the results of the present studies are considered indicative of reliable characteristics of the vicarious emotional responder (and the replication and extension provided by Study 2 clearly support this), a methodological dilemma may have confounded the results. That is, the absence of a control group that experienced the guest speaker "in his usual form" prevented our conclusions from being clear-cut. The speaker himself and the manipulation were confounded. Our results could have been due to "the nature of our guest speaker." Analysis of changes in state anxiety from a nonstress to a post-guest-speaker period during which the guest speaker did not perform the anxiety manipulation was a needed control. The purpose of Study 3 was to examine such a control by analyzing state anxiety changes before and after the same guest speaker for subjects varying in empathy and trait anxiety.

Method

Subjects

The subjects in this study were 14 undergraduate psychology students. There were 2 males and 12 females, with a mean age of 23.2. All were students in a low-enrollment, summer-session undergraduate psychology class at an urban Virginia university.

Procedure

In this study the STAI and the empathy scale were administered to the entire class during a nonstress condition. The second administration of A-State followed a guest talk on the same topic as in Studies 1 and 2 and by the same guest speaker. In this study the speaker did not perform the anxiety manipulation.

Results and Discussion

Empathy

Empathy was found to be nonsignificantly related to both the initial A-State ($r = -.18$) and A-State increases ($r = -.08$). Even though both Study 1 and 2 found empathy to be significantly related to A-State increases, there was no such relationship in this study when the vicarious anxiety experience was not provided.

When subjects were divided at the median of Studies 1 and 2 (38) into high-empathy

(>38, $n = 9$) and low-empathy (>38, $n = 5$) groups, it was possible to match 5 subjects on their initial A-State scores. The means for the high- and low-empathy subjects that had been matched on initial A-State were pre = 34.4 and 34.4 and post = 32.2 and 33.6, respectively. The results of the analysis of variance indicated nonsignificant main effects for empathy and trials and a nonsignificant interaction, $F_s(1, 12) < 1$. These results indicate that with matched subjects⁶ there is no difference in A-State for subjects varying in empathy and that the guest speaker did not produce anxiety arousal. Thus, without the anxiety manipulation the speaker did not arouse anxiety.

Anxiety

Trait anxiety was significantly related to initial A-State scores ($r = .64$, $p < .05$) but was not significantly related to state anxiety increases ($r = .16$). As expected, unlike Studies 1 and 2 but in agreement with the empathy data from Study 3, A-Trait was not related to A-State increases.

When subjects were divided at the median (38) of their A-Trait scores into a high A-Trait (>38, $n = 7$) and low A-Trait (<38, $n = 7$) groups, the results of a 2 (A-Trait level) \times 2 (trials) analysis of variance indicated a significant main effect for A-Trait level, $F(1, 12) = 5.38$, $p < .05$, a nonsignificant main effect for trials, and a nonsignificant interaction, $F_s(1, 12) < 1$. The means for high and low A-Trait subjects were pre = 39.86 and 32.43 and post = 38.00 and 31.86. These results support previous findings of higher A-State for high A-Trait subjects. Also, these results indicate that the guest speaker (without the anxiety manipulation) did not produce elevations in state anxiety. In addition, there was no differential responsiveness for high versus low A-Trait subjects.

The results of the control study demonstrated that the observed anxiety reactions in Studies 1 and 2 were probably a function of the anxiety manipulation and not due to

⁶ Similarly nonsignificant results were obtained when the median split groups were analyzed.

any idiosyncrasies of the speaker himself. Thus, those findings reported in Study 1 and Study 2 should not be considered subject to a confound due to the speaker's style.

Reference Note

1. Kendall, P. C., Finch, A. J., Jr., & Mikulka, P. J. *Multidimensional locus of control in adults*. Unpublished manuscript, Virginia Commonwealth University, 1976.

References

- Abramowitz, S. I. Internal-external control and social-political activism: A test of the dimensionality of Rotter's internal-external scale. *Journal of Consulting and Clinical Psychology*, 1973, 40, 196-201.
- Alfert, E. Comparison of responses to a vicarious and a direct threat. *Journal of Experimental Research in Personality*, 1966, 1, 179-186.
- Alfert, E. An ideographic analysis of personality differences between reactors to a vicarious experienced threat and reactors to a direct threat. *Journal of Experimental Research in Personality*, 1967, 2, 200-207.
- Auerbach, S. M. Effects of orienting instructions, feedback-information, and trait-anxiety level on state anxiety. *Psychological Reports*, 1973, 33, 779-786.
- Averill, J. R., Olbrich, E., & Lazarus, R. S. Personality correlates of differential responsiveness to direct and vicarious threat: A failure to replicate previous findings. *Journal of Personality and Social Psychology*, 1972, 21, 25-29.
- Bandura, M. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Bennett, D. H., & Holmes, D. S. Influence of denial (situation redefinition) and projection on anxiety associated with threat to self-esteem. *Journal of Personality and Social Psychology*, 1975, 32, 915-921.
- Endler, N. S., & Okada, M. A multidimensional measure of trait anxiety: The S-R Inventory of General Trait Anxiousness. *Journal of Consulting and Clinical Psychology*, 1975, 43, 319-329.
- Gleser, G. C., & Ihilevich, D. An objective instrument for measuring defense mechanisms. *Journal of Consulting and Clinical Psychology*, 1969, 33, 51-60.
- Gleser, G. C., & Sacks, M. Ego defenses and reaction to stress: A validation study of the Defense Mechanisms Inventory. *Journal of Consulting and Clinical Psychology*, 1973, 40, 181-187.
- Hodges, W. F., & Spielberger, C. D. The effects of threat of shock on heart rate for subjects who differ in manifest anxiety and fear of shock. *Psychophysiology*, 1966, 2, 287-294.
- Hogan, R. Development of an empathy scale. *Journal of Consulting and Clinical Psychology*, 1969, 33, 307-316.
- Holmes, D. S., & Houston, B. K. The defensive function of projection. *Journal of Personality and Social Psychology*, 1971, 20, 208-213.
- Holmes, D. S., & Houston, B. K. Effectiveness of situation redefinition and affective isolation in coping with stress. *Journal of Personality and Social Psychology*, 1974, 29, 212-218.
- Houston, B. K. Trait and situational denial and performance under stress. *Journal of Personality and Social Psychology*, 1971, 18, 289-293.
- Houston, B. K. Viability of coping strategies, denial, and response to stress. *Journal of Personality*, 1973, 41, 50-58.
- Houston, B. K., & Hodges, W. F. Situational denial and performance under stress. *Journal of Personality and Social Psychology*, 1970, 16, 726-730.
- Kendall, P. C. Anxiety: States, traits—Situations? *Journal of Consulting and Clinical Psychology* 1978, 46, 280-287.
- Kendall, P. C., Finch, A. J., Jr., Auerbach, S. M., Hooke, J. F., & Mikulka, P. J. The State-Trait Anxiety Inventory: A systematic evaluation. *Journal of Consulting and Clinical Psychology*, 1976, 44, 406-412.
- Kendall, P. C., Finch, A. J., Jr., & Mahoney, J. Factor specific differences in locus of control for emotionally disturbed and normal children. *Journal of Personality Assessment*, 1976, 40, 42-45.
- Lazarus, R. S. *Psychological stress and the coping process*. New York: McGraw-Hill, 1966.
- Lazarus, R. S., & Alfert, E. Short-circuiting of threat by experimentally altering cognitive appraisal. *Journal of Abnormal and Social Psychology*, 1964, 69, 195-205.
- Lazarus, R. S., Opton, E., Nomikos, M., & Rankin, N. The principle of short-circuiting of threat: Further evidence. *Journal of Personality*, 1965, 33, 622-635.
- Levenson, H. Multidimensional locus of control in psychiatric patients. *Journal of Consulting and Clinical Psychology*, 1973, 41, 397-404.
- Nowicki, S., & Duke, M. P. A locus of control scale for noncollege as well as college students. *Journal of Personality Assessment*, 1974, 38, 136-137.
- Nowicki, S., & Strickland, B. A locus of control scale for children. *Journal of Consulting and Clinical Psychology*, 1973, 40, 143-154.
- Opton, E., & Lazarus, R. Personality determinants of psychophysiological response to stress: A theoretical analysis and an experiment. *Journal of Personality and Social Psychology*, 1967, 6, 291-303.
- Phares, E. J. *Locus of control: A personality determinant of behavior*. Morristown, N.J.: General Learning Press, 1973.
- Ray, W. J., & Katahn, M. Relation of anxiety to locus of control. *Psychological Reports*, 1968, 23, 1196.
- Sarason, I. G. Experimental approaches to test anxiety: Attention and the use of information. In C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (Vol. 2). New York: Academic Press, 1972.
- Speisman, J. C., Lazarus, R. S., Mordkoff, A., & Davison, L. Experimental reduction of stress based on ego-defense theory. *Journal of Abnormal and Social Psychology*, 1964, 68, 367-380.
- Spielberger, C. D. Anxiety as an emotional state. In

- C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (Vol. 1). New York: Academic Press, 1972.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *Manual for the State-Trait Anxiety Inventory* (self-evaluation questionnaire). Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Stevens, H. A., & Reitz, W. E. An experimental investigation of projection as a defense mechanism. *Journal of Clinical Psychology*, 1970, 26, 152-154.
- Watson, D. Relationship between locus of control and anxiety. *Journal of Personality and Social Psychology*, 1967, 6, 91-92.
- Zuckerman, M. Development of a situation-specific trait-state test for the prediction and measurement of effective responses. *Journal of Consulting and Clinical Psychology*, 1977, 45, 513-523.

Received July 26, 1977 ■

Illusory Correlation: A Further Exploration of Chapman's Paradigm

Richard M. Kurtz and Sol L. Garfield
Washington University

Even though the phenomenon of illusory correlation first researched by Chapman has been demonstrated in a variety of studies, relatively little has been done to show convincingly that this bias cannot be reduced or eliminated by training. The present study addresses itself to this issue. There were four groups of 15 undergraduate subjects each in this study. The first group was a replication of a study by Chapman and Chapman with an equal association of all valid Wheeler signs and invalid signs and statements of the patients' purported problem. Group 2 was a replication of Chapman's study, with valid signs presented 100% of the time and invalid signs presented 50% of the time. Group 3 provided subjects with a special pretraining against illusory correlation, with 50% presentation of valid signs; and Group 4 was a special pretraining of the subjects against illusory correlations, with 100% presentation of the valid Wheeler signs. It was predicted that Groups 3 and 4 would show the least amount of illusory correlation. This hypothesis was not confirmed. However, we replicated part of the Chapmans' findings that subjects predominantly associated the concept of anality with preconceived problems of homosexuality. A serendipitous finding was also noted in which subjects appeared to create their own illusory correlate.

Illusory correlations are reported associations between test responses and symptoms or syndromes that are based on verbal associative connections of the test-sign to the symptom rather than on valid observations. As defined by Chapman (1967), illusory correlation is

the report by observers of the correlation between two classes of events which, in reality, (a) are not correlated, or, (b) are correlated to a lesser extent than reported, or (c) are correlated in the opposite direction from that which is reported (p. 151).

Although the phenomenon of illusory correlation has been demonstrated in a variety of studies (Chapman, 1967; Chapman & Chapman, 1967, 1969; Golding & Rorer, 1972; Starr & Katkin, 1969), relatively little has been done to show that this bias cannot be

reduced or eliminated by proper training. The present study addresses itself to that issue.

The phenomenon of illusory correlation was first demonstrated by Chapman (1967) when subjects observed a random pairing between elements of two arrays of words and were then asked to report the frequency with which a word from one array was paired with a word from another array. He found the reported frequency of occurrence was biased upwards when the word pair was characterized by either strong verbal associative connections or by distinctiveness.

In a second study (Chapman & Chapman, 1967), subjects were shown randomly paired symptom statements and Draw-a-Person figures. They were then asked to report any drawing characteristic that they felt was associated with any of the symptoms. The relationships that the subjects reported tended to be similar to those that some clinicians reported from their clinical practices, and they basically seemed to rely on associative

Requests for reprints should be sent to Richard M. Kurtz, Department of Psychology, Washington University, St. Louis, Missouri, 63130.

connections between the drawing characteristics and the symptoms.

In one of the nuclear studies conducted by Chapman and Chapman (1969) on the Rorschach, they found that expert diagnosticians, on the basis of their clinical experience, and naive judges, on the basis of their observations of random materials, tended to report the same Rorschach signs as being valid indicators of homosexuality. This was found to be the case even though these signs had not been demonstrated in the previous research literature to relate to the purported problem of homosexuality, nor did they have, in the experimental task, a nonrandom association with the purported problem of homosexuality. The basic experimental task used by Chapman and Chapman was to present a series of Rorschach cards in which for each Rorschach response or percept two statements of an emotional problem allegedly given by a purported patient were paired to that particular part of the Rorschach blot. The task of the subjects in the Chapman study was to look at these paired associates of the Rorschach percept and the statement of the problem and, at the end, to indicate whether or not they noticed any kind of response that was seen most often by patients with this particular problem.

Consistent with previous studies, Chapman and Chapman (1969) found that subjects picked those percepts that had the highest verbal associative connection with the symptom, in this case homosexuality. In addition, Chapman and Chapman (1969) demonstrated that subjects still tended to report the popular invalid sign¹ disproportionately even when it was paired randomly with the symptom "homosexuality," and the unpopular valid sign was paired 100% of the time with the same symptom.

In an extensive replication, Golding and Rorer (1972) came to the same conclusion. They found in replicating Chapman and Chapman's (1969) study on illusory correlation that little change occurred even when the nonillusory (Chapmans' so-called "valid") cues were paired 100% of the time with the symptom of homosexuality and when the illusory cues (Chapmans' so-called "invalid" signs) had a randomly paired relationship

with the symptom of homosexuality. They also found that the different modes of feedback to the subjects and different symptom-base rates did not produce differential effects in posttraining estimates of the illusory correlation. From this they concluded that training (as experimentally defined) has very little effect on the illusory correlation.

The present study attempts to explore whether training can effect the illusory correlation. One of the problems inherent in the previous research is that attempts to modify the illusory correlation have only been minimally manipulated with the device of changing the amount or the percentage of time that nonillusory (the so-called valid) or the illusory signs (the so-called invalid) are paired with the purported problem of homosexuality in the experimental task. Perhaps with more explicit training, that is, warning the subjects that they must be on guard against illusory correlations, and showing them examples of such associative connections, that perhaps the subjects' tendency to fall prey to illusory correlations could be reduced.

There were four groups in the present study. The first group was a replication of a part of Chapman and Chapman's (1969) study (Experiment 2, p. 275) with an equal association, that is 50% presentation of all valid and invalid signs with all statements of the patient's purported problem. The valid signs were those designated by Chapman and Chapman (1969) based on the work of Wheeler (1949). Group 2 was a replication of a part of Chapman and Chapman's (1969) study (Experiment 3, p. 277), with valid signs presented

¹ For detail of the discussion of how Chapman and Chapman (1969) empirically developed the concept of valid and invalid signs, see their article (pp. 272-275). As the methodology was quite complex and space is limited, suffice it to say that a "valid sign" was one that they found in the literature to be empirically related to actual reported homosexuality, and "invalid signs" were ones that clinicians believed predicted homosexuality but in fact did not. The valid and invalid signs were either randomly assigned to a variety of different purported problems or were presented either 50% or 100% of the time with the actual problem of homosexuality. The present study duplicates part of their total experiment, that is, their Experiments 2 and 3.

100% of the time with the symptom of homosexuality and invalid signs presented 50% of the time with the symptom of homosexuality. Group 3 was a special pretraining of the subjects against illusory correlation with 50% presentation of valid signs, and Group 4 was a special pretraining of the subjects against illusory correlation with 100% presentation of valid signs. It was predicted that Groups 3 and 4 would show the least amount of illusory correlation. In other words, a linear trend was expected, with Group 1 showing the greatest amount of illusory correlation, then Group 2, Group 3, and Group 4 showing the least.

Method

Subjects

Four groups of 15 volunteer subjects of both sexes² each ($N=60$) were secured from undergraduate abnormal psychology classes at Washington University. The subjects were told that the authors were studying training parameters that affect Rorschach interpretation and that they needed the help of an untrained population to study basic patterns of test interpretation.

Test Instrument

The same materials used by Chapman and Chapman (1969) were used in the present study. These consisted of 30 Rorschach cards that were transferred to transparencies. On each of the transparencies, one percept or response was paired with two statements of emotional problems of a purported patient who was alleged to have given that specific response. The Rorschach percepts were indicated by circling an area of the card and pasting a typed statement of the patient's verbalization. For example, one of the 30 Rorschach responses labeled *Bugs Bunny* was given to the center area (D7) of card 5 (Beck, Beck, Levitt, & Molish, 1961). In the corner of the card appeared the statement, "The man who said this (1) had sexual feelings toward other men and (2) feels sad and depressed much of the time." The 30 percepts were chosen so that 6 fell into each of the five following categories: (a) popular invalid sign or a sign highly connected with associative material, in this case, human or animal anal content; (b) Wheeler Sign No. 7 (for more complete discussion of the Wheeler signs, see Chapman & Chapman, 1969); (c) Wheeler Sign No. 8; (d) geographical features, a filler category; and (e) food, another filler category.

The two statements of emotional problems or symptoms listed on the cards were drawn from a pool of four such statements, and these were identical to those used by Chapman and Chapman (1969).

They were (a) "he has sexual feelings toward other men"; (b) "he believes other men are plotting against him"; (c) "he feels sad and depressed much of the time"; and (d) "he has strong feelings of inferiority." The statements of symptoms (diagnostic cues) and Rorschach percepts were paired on the 30 cards so that each of the four statements appeared 15 times in Groups 1 and 3. Each symptom statement was thus paired with 3 of the 6 percepts from each of the five categories of percepts. Thus, for Groups 1 and 3 there was no intrinsic relationship between the occurrence of any one of the four symptoms and any one of the five categories of response. In Groups 2 and 4, the valid statement (both Wheeler signs) occurred 100% of the time with the purported problem of homosexuality, and the invalid statements occurred 50% of the time.

Procedure

The subjects were given some brief introductory information as to the nature of the Rorschach. However, no information was given about categories of either content or determinants. All groups received the following instructions:

I am going to give you a brief introduction to the well-known Rorschach test and then ask you to perform some exercises on it.

As many of you know, the Rorschach is a test of personality functioning. It basically consists of a set of inkblots to which the subject is asked to respond by describing what he perceived in each of the inkblots. It is believed that what the individual sees and describes will reveal important aspects of his personality. If he is concerned or fearful about certain problems or has conflicts in certain areas, these may be reflected in his responses to the Rorschach inkblots. How the individual perceives different blots or segments of the blots, as well as the content of his responses, may thus tell us some potentially important things about his personality.

This will suffice as a brief, general overview of what the test aims to appraise.

I am going to show you a series of inkblots, one at a time. On each inkblot you will find a typed statement of what one patient saw on this blot and also what his two chief emotional problems are. Each of these 30 cards represents a different patient. You will see what 30 different patients

² Chapman and Chapman, in their 1969 study, did not designate the sex of the subjects. In the present study, half of the subjects were male and half were female, and they were randomly assigned to each of the four training conditions. No further analysis was done by sex.

said they saw on a card. Now let me tell you what I want you to do. Please carefully study each ink-blot and the statement of what the patient said he saw in it. Also, study the statement of the patient's two severe emotional problems. When everyone has looked at all of the cards, I'm going to give you a questionnaire in which I will ask you about the kinds of things seen by patients with each kind of problem. [This paragraph is from Chapman and Chapman, 1969, p. 276]. Is that clear? All right, then, we'll go ahead.

Besides the above instructions Groups 3 and 4 also received these additional training instructions:

Before we proceed, however, it is worthwhile to also present a few important cautions. Sometimes interpretations are made that are not based on actual empirical findings about the test. Instead, individuals may make interpretations which are based on what *appear to be* verbal or logical similarities between the content of the response and a certain type of personality problem. Let me give you some examples of what I mean. For example, since paranoid individuals are seen as suspicious, responses which depict "eyes" or "someone staring" may be interpreted as indications of paranoid personality—even though this has not been really demonstrated. In a similar fashion, psychologists have interpreted responses depicting explosions or destruction as signifying aggression on the part of the person making such responses. In other words, these individuals are making some associational tie between the response and the interpretation—although such interpretations have not been supported by subsequent research. This is a common error in Rorschach interpretations and should be guarded against. You should be on your guard and alerted against making such obvious kinds of associations which seem to go together verbally but actually are not related. Instead, try to base your conclusions on the personality characteristics which have been found to be associated with certain kinds of responses to the ink blots in the material to be presented to you. Since this is a very important aspect, let us go over these points again. [This paragraph was then repeated.]

The cards were then circulated in a prearranged pattern so that each subject saw each of the 30 cards for 60 sec. The order of presentation was systematically counterbalanced so that none of the content categories (i.e., Wheeler 8, Wheeler 7, geography, anal, food) appeared more than once in a sequence. After the 30 cards were presented, the subjects were given the following questionnaire, which is identical to that used by Chapman and Chapman (1969, p. 276).

Some of the things in inkblots were seen by men who had the following problem:

He had sexual feelings toward other men.

Did you notice any kind of general thing that was seen most often by men with this problem?

Yes _____ No _____

If your answer is yes, name that kind of thing and give one example of that kind of thing. Kind of thing _____

Example _____

The identical format was followed for the other three diagnostic problems: "He has strong feelings of inferiority"; "he feels sad and depressed much of the time"; "he believes other people are plotting against him." After the subjects completed this questionnaire, they were asked to list any and all of the reactions and hypotheses that they formulated while participating in the experiment. Following this, they were debriefed.

Results

To test the major hypothesis that special training could affect the illusory correlation, a 4×4 repeated measures analysis of variance was performed on dichotomous data (Winer, 1962) with repeated measures on the second factor. The mean number of illusory correlates seen by Group 1 (50% presentation valid-invalid signs; no training) was 7.8; the mean for Group 2 (100 valid, 50% invalid; no training) was 8.0; for Group 3 (50% presentation valid-invalid; training) was 9.5; and the mean for Group 4 (100% presentation valid, 50% invalid; training) was 9.3.

The main effects for the training groups were insignificant, ($F < 1$; MS for treatments = .21, $df = 3$; error = .31, $df = 56$). A significant main effect for diagnostic cue was found, $F = 19.1$, MS for cue = 3.4, $df = 3$; MS error = .18, $df = 168$). The interaction was also insignificant ($F < 1$).

The mean number of illusory correlates seen for the cue of sex was 11.5; for inferiority, $M = 10.5$; for depression, $M = 3.5$, and for paranoid, $M = 9.0$, with the difference between the means of depression and the other three accounting for the significant F value.

Examining the data in the same way that Chapman and Chapman (1969) did, a chi-square was computed on the sexual variable of homosexuality and broken into four categories—(a) number of anal responses seen; (b) number of combined valid Wheeler signs seen; (c) number of filler items, that is.

Table 1
Percentage of Illusory Correlates Given by All Subjects to Diagnostic Cues

Response	Purported patient problems (diagnostic cues)			
	Inferiority	Sexual	Depression	Paranoid
Anal (invalid sign)	2	55	3	0
Wheeler (valid sign)	13	20	2	23
Geography	5	2	5	20
Food	3	0	5	3
Other	2	0	5	3
"Looks small"	45	0	3	12
No response	30	23	77	40

geography, food, other; and (d) no illusory correlation seen. Since only one response occurred in Category 3, this category was dropped and the response was placed in Category 4. The obtained frequencies for Category 1 was 33; for Category 2, 12; and for Collapsed Categories 3 and 4, 15. The expected frequencies were 20 in each category. The obtained chi-square of 12.9 ($df = 2$) was significant beyond the .005 level and was consistent with the data reported by Chapman. Thus, in this part of the study we completely replicated part of the findings of Chapman and Chapman (1969, Experiments 2 and 3).

Due to the unexpectedly high number of illusory correlates seen for the category of inferiority as well as paranoia, all data were cast into Table 1, which shows the percentage of illusory correlates for the purported problems of patients and the content of the Rorschach area. Since the initial category of "other" was composed predominantly of subjects' statements that said that the patient had problems of inferiority because the area seen by the patient was small or tiny, this category was then divided into two: "other" and "looks small." Thus, a post hoc category of "things that look small"^a was added. The remainder of the other category for problems of inferiority accounted for only 2% of the illusory correlation, whereas the new category of "seeing things as small" accounted for 45%. On the other hand, the expected anal content far outweighed that of the Wheeler signs for the problem of homosexuality. Fifty-five percent of the subjects gave the response "anal content" for what they most often saw from men with sexual

problems. None of the subjects used the category looks small in association with sexual problems. There does not appear to be any consistent pattern with the purported problem of paranoia, as 20% of the subjects used Wheeler signs, 20% used geography, and 12% used category looks small. The largest majority, 77%, saw no illusory correlates in relationship to the problem of depression.

The frequency of percepts being characterized as looks small suggests that the subjects may have invented an illusory correlate for the problem of inferiority. If this were so, it would be significant, since this concept was not linguistically built into the experimental material.

Even though the data are convincing for Chapman's position, when one examines the sex problems versus the Wheeler signs, it appears from the overall data that the phenomena of illusory correlation may be much more complex. Besides the area of content-triggered interpretations, it would appear that subjects try to find some meaning in the clinical material presented. In the case of inferiority, the subjects seem to have selected an aspect of the blot configuration associated with smallness and then associated this aspect with "inferiority feelings" that were not, in a strict sense, linguistically triggered by the

^a The category "seeing things as small" consisted of the subjects' responses to small details of the Rorschach blot (see Beck, Beck, Levitt, & Molish, 1961) and statements of "small in area," "small things," "tiny things," and so on. A response was not considered belonging to this category if any other content was indicated, that is, "small island."

content. Also, considering the high number of illusory correlates (i.e., 60%) given by subjects to the problem of paranoia, it does appear as if there is a strong tendency among subjects to find meaning in clinical material even if such meaning does not exist.

Discussion

In spite of the fact that attempts were made to influence the illusory correlation by providing a simulated training session for the subjects, it was not possible to reduce the illusory correlation. Under either the 100% or the 50% presentation conditions, training was no more effective than nontraining conditions in reducing the effect of this phenomenon. This finding is consistent with that of Chapman and Chapman (1969) and Golding and Rorer (1972). It must be recognized, however, that the training conditions in this study and in the other two just mentioned represent a limited attempt to modify a phenomenon that appears to be basic in the way people conceptualize diagnostic problems. Whether more intensive training with constant feedback and with numerous examples of faulty associational thinking would reduce the illusory correlation is a question for future research. This study supports the Chapman and Chapman (1969) position that the illusory correlation is a robust phenomenon.

Although the training conditions were not effective in reducing the illusory correlation, we replicated part of the Chapman and Chapman (1967) findings that the concept of "anality" was predominantly associated with preconceived problems of homosexuality regardless of training conditions and frequency of presentation of valid signs. A serendipitous finding was also noted in which the subjects appeared to create their own illusory correlate. That is, they created a concept of seeing "small things" and associated it significantly with problems of inferiority. Since there was no linguistic content in the stimulus material per se, it can be hypothesized that this represents a structural illusory correlate, that is, small segments of the blot rather than specific content as such. In other words, illusory correlations may span more than just simple linguistic associations. This

is an area that needs to be explored in further research.

At the end of the experiment, all of the subjects were asked for their spontaneous comments, ideas, and expectations as to what was being tested. These were recorded, and, although they do not allow any systematic quantitative analysis, they reveal that the majority of subjects expressed strong doubts as to whether the Rorschach had any valid meaning. In spite of this skepticism on the part of the subjects, the majority of them formed illusory correlations. The subjects also complained about the nature of this task, stating that it was excessively long and boring. For Groups 2 and 4 the creation of nonrandom contingencies for the Wheeler signs means that nonrandom contingencies were created for the other three categories. This experimental format, of course, is identical to that used by Chapman and Chapman. None of the subjects, however, acknowledged this either in their responses to the valid Wheeler signs or in spontaneous comments in the debriefing.

Another possible factor is the matter of potential response biases that may be imbedded in the form of the questionnaire used by Chapman and Chapman. For instance, Chapman and Chapman's (1969) questionnaire asked: "Did you notice any kind of general thing that was seen most often by men with this problem? If your answer is 'yes,' name that kind of thing and give *one* example of that kind of thing." Such instructions clearly establish a strong demand characteristic for subjects not only to find something but to pick the most easily remembered, or salient, feature. Percepts like "anus" clearly fall into this class. However, Golding and Rorer (1972) used a form of prediction feedback in their study that was expected to lead to a greater reduction in the amount of illusory correlation, but they found that the illusory correlation remained a robust one.

Also neglected in the research is the question of social pressures; that is, all subjects were tested in groups. Perhaps some felt that since others were writing down their comments, they should also be finding things, even though they themselves did not believe that the task was meaningful. In other words,

the questionnaire asked whether people had seen something, but the form of the questionnaire seems to have been interpreted by the subjects as a demand to find meaning in something that they felt was predominantly puzzling. Future research should investigate this by modifying the form or by leaving the task response more open-ended. Perhaps, providing the subjects with a range of behavioral statements about people with so-called problems, allowing them to use an actuarially designed cue sort, or allowing them to describe what they thought the predominant problems of these subjects were would produce a significantly different picture with regard to illusory correlates. With these modifications, subjects conceivably might rely less on associationally triggered responses that are paired with the diagnostic symptoms in a highly artificial, experimental paradigm.

Finally, it can be noted that the experimental arrangement developed by Chapman and Chapman (1969) does not closely approximate the traditional diagnostic process of trained clinicians and may limit the generalizability of these studies to clinical situations. Additional research might focus on attempts to do more in vivo studies of clinicians' actual conceptualizations and investigate whether the justification for the clinical

inferences can be reduced primarily to associational variables or whether in fact they represent rather illogical or irrational learning or teaching that clinicians have incorporated during their graduate education.

References

- Beck, S. J., Beck, A. G., Levitt, E. E., & Molish, H. B. *Rorschach's test*. New York: Grune & Stratton, 1961.
- Chapman, L. Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior*, 1967, 6, 151-155.
- Chapman, L., & Chapman, J. Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, 1967, 72, 193-204.
- Chapman, L., & Chapman, J. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 1969, 74, 271-280.
- Golding, S. G., & Rorer, L. Illusory correlation and subjective judgment. *Journal of Abnormal Psychology*, 1972, 80, 249-260.
- Starr, J. G., & Katkin, E. The clinician as an aberrant actuary: Illusory correlation and the Incomplete Sentences Blank. *Journal of Abnormal Psychology*, 1969, 74, 670-675.
- Wheeler, W. M. An analysis of Rorschach indices of male homosexuality. *Rorschach Research Exchange*, 1949, 13, 97-126.
- Winer, B. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Received July 26, 1977 ■

Avoidance-Approach: The Fifth Basic Conflict

Seymour Epstein

University of Massachusetts at Amherst

The basic conflicts are almost always listed as approach-approach, approach-avoidance, avoidance-avoidance, and double approach-avoidance conflict. The possibility of avoidance-approach conflict, in which a steeper gradient of approach intersects a gradient of avoidance, is ignored because it is assumed that approach gradients cannot be steeper than avoidance gradients, and that even if avoidance-approach conflict could exist, it would be of no interest, as the individual would simply stay away from the conflicting goal. It is demonstrated that there is no good reason for assuming that approach gradients cannot be steeper than avoidance gradients, and there is considerable evidence that they often are. It is further noted that individuals can be placed in situations not of their choosing. If an individual with an avoidance-approach conflict were placed on the goal side of the intersection of the gradients, the person would enthusiastically approach a goal that had previously been avoided. Thus, avoidance-approach conflict can account for ego-alien behavior, such as when a shy, sex-avoidant "model" boy commits a violent crime of passion. The implications and causes of avoidance-approach conflict in everyday life are discussed.

In almost all textbooks of introductory and abnormal psychology, four basic conflicts are listed with accompanying diagrams. These are approach-approach, avoidance-avoidance, approach-avoidance, and double approach-avoidance conflict. Particular attention is then given to approach-avoidance conflict, because it can account for a variety of clinical phenomena, such as that an individual in such a conflict tends to become entrapped at a midpoint from a goal and to suffer, as a consequence, from a state of heightened drive. The entrapment is explained by the assumption that the gradient of avoidance is steeper than the gradient of approach. No consideration is given to the possibility of a conflict, which I shall refer to as "avoidance-approach" conflict, in which a steeper gradient of approach intersects a gradient of avoid-

ance. There are probably two reasons why avoidance-approach conflict has been ignored. One is that it is assumed that the gradient of avoidance is necessarily steeper than the gradient of approach. The other is that it is assumed that even if a case could be made for such conflict on theoretical grounds, it would be of no practical interest, as an individual with a flatter gradient of avoidance than of approach would simply stay away from the conflicted goal. It will be demonstrated in this article that approach gradients can be steeper than avoidance gradients, and that there are important clinical phenomena, such as ego-alien behavior, that can be accounted for once this is recognized.

Lewin (1935), who originally formulated the concept of approach-avoidance conflict, simply stated without evidence that "the negative vector usually increases gradually in strength and finally becomes stronger than the positive" (p. 90). It is not clear whether he meant the word *usually* to indicate that given a negative incentive, there is usually a goal gradient, which would appear to be self-evident, or whether it was to indicate that the negative vector usually overtakes the posi-

This article was supported by National Institute of Mental Health Research Grant MH-01293.

I wish to express my appreciation to James Averill for his helpful criticism of an earlier version of this article.

Requests for reprints should be sent to Seymour Epstein, Department of Psychology, University of Massachusetts, Amherst, Massachusetts 01003.

tive. In the latter case, it would indicate that Lewin did not believe that the positive vector is always steeper than the negative one, but when it is, some interesting consequences follow.

Miller (1944) also initially assumed without explanation that avoidance gradients are steeper than approach gradients. More recently, Miller (1959) revised his position and noted that there is no intrinsic reason for avoidance gradients to be steeper than approach gradients, and that whether they are steeper or less steep depends on the extent to which the gradients are based on inner relative to outer cues. He observed that in many of the studies in which steeper gradients of avoidance were found, the avoidance gradients were based on the delivery of punishment, an external source of stimulation, whereas the approach gradients were based on hunger, an internal source of stimulation. It appears that there is no logical basis for assuming that approach gradients cannot be steeper than avoidance gradients.

As for empirical evidence on steepness of approach and avoidance gradients, it has been demonstrated that a number of factors (cf. review in Heilizer, 1977), such as stimulus differentiation (e.g., Bugelski & Woodward, 1951; Hearst, 1962; Saltz, Whitman, & Paul, 1963), number of trials (e.g., Elder, Kuehne, Clarke, & Larre, 1970; Elder, Kuehne, & Moriarty, 1970; Schroeder & Gerjuoy, 1965; Weiss, 1960), runway length (e.g., Clifford, 1973), and mental age (Tempone, 1965), are directly related to steepness of gradients. Among these, stimulus differentiation appears to be the most important factor, as it can account for the others. Thus, the finding that runway length is an important factor can be accounted for by the consideration that the greater the runway length, the more easily the end and the beginning of the runway can be differentiated (Saltz et al., 1963). The influence of number of reinforced trials can be attributed to an increase in stimulus discrimination that occurs over trials, and the influence of mental age can be attributed to the more accurate discrimination of children with higher mental age. Stimulus differentiation can also account for Miller's observation that avoidance gradients have

been found to be steeper than approach gradients when avoidance is contingent on external cues and approach on internal cues, as internal cues may not be as easily discriminated as external cues.

It follows from the above that depending on such factors as training in stimulus discrimination, number of reinforced trials, and degree to which motives are based on external relative to internal stimuli, approach gradients can be less steep, more steep, or no different in steepness from avoidance gradients. This conclusion is supported by a number of studies that have directly compared the steepness of approach and avoidance gradients. Some have reported avoidance gradients to be steeper than approach gradients (e.g., Brown, 1948; Miller & Kraeling, 1952; Miller & Murray, 1952; Murray & Berkun, 1955), some have reported approach gradients to be steeper than avoidance gradients (e.g., Hearst, 1960, 1962; Smith, 1965, 1969), and some have reported no difference in steepness of approach and avoidance gradients (e.g., Desiderato, Foldes, & Gockley, 1966; Gjesme, 1974; Hearst, 1960; Rigby, 1954).

Following an analysis of the relative steepness of approach and avoidance gradients, Maher (1966) noted that there is no firm support for the assumption that avoidance gradients tend to be steeper than approach gradients. He observed that the classic studies of Brown (1948), widely cited as evidence for the assumption, have serious methodological flaws. In a study by Maher and Nuttall (in Maher, 1966), order of testing was found to influence the relative steepness of gradients, and Maher concluded that Brown's finding of a steeper gradient of avoidance than of approach could be attributed to fatigue effects associated with order of testing. Even in studies that report no significant difference in steepness of avoidance and approach gradients, it should be considered that some subjects produced steeper gradients of approach and others of avoidance. That is, there are individual differences with respect to relative steepness of approach and avoidance gradients. In a study by Gewirtz (1959) that directly dealt with this issue, reliable individual differences were found in the relative steepness of approach and avoidance gradients. In a

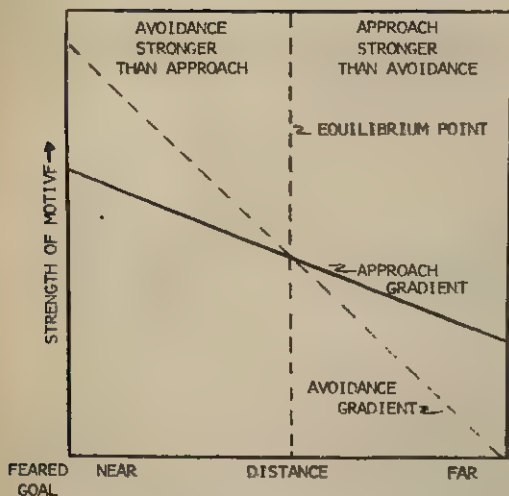


Figure 1. Approach-avoidance conflict. (The individual approaches the goal when at a distance, vacillates at the intersection of the gradients, and avoids the goal when closer to it. The result is that the individual becomes trapped by his or her own motives, neither being able to fulfill nor abandon the approach motive.)

recent study in our own laboratory (Losco & Epstein, 1977), relative steepness of avoidance and approach gradients was investigated while holding magnitude and quality of incentive constant. Although a weak, significant tendency was found for the avoidance gradient to be steeper than the approach gradient, inspection of the results for individuals revealed that the phenomenon was far from uniform among subjects, with some producing much steeper approach than avoidance gradients. It was further noted that the mean tendency was not very robust and was readily canceled out by incidental factors, such as motor exertion.

Considering all the evidence together, it can safely be concluded that it is possible for approach gradients to be steeper than avoidance gradients. As already noted, Miller has come to the same conclusion. He has indicated that the only reason that he did not pursue its implications is that he found a steeper gradient of approach than of avoidance to be "less perspicuous than the other pattern" (Miller, 1959, p. 222).

A pattern in which a steeper gradient of approach intersects a gradient of avoidance not only has significant implications in its own

right, but it is important to recognize it in order to avoid confusion between approach-avoidance and avoidance-approach conflict. It is instructive, in this respect, to contrast the major features of the two types of conflict. There are three major behavioral consequences for an individual with an approach-avoidance conflict. A person with such a conflict approaches the conflicted goal at a distance, is blocked and vacillates at an intermediate point at which the gradients intersect, and retreats when closer to the goal (see Figure 1). According to Miller, the most efficacious treatment for approach-avoidance conflict is to reduce the avoidance drive, as attempting to force the individual to reach the goal can induce unmanageable levels of anxiety, and would very likely fail, since avoidance tendencies might well mount more rapidly than the combined approach tendencies. Avoidance-approach conflict also has three behavioral consequences. The individual in an avoidance-approach conflict avoids the goal when on the far side of the intersection of the gradients, is hesitant and blocked, at least momentarily, when at the point of intersection of the gradients, and exhibits an accelerated approach response if placed on the goal side of the intersection point (see Figure 2). Thus, in both approach-avoidance and avoidance-approach conflict, there are manifestations of strong avoidance reactions and of blocking and hesitation, although in the case of avoidance-approach conflict the blocking is of short duration, as movement in either direction breaks the deadlock. Considering that what is near and far from a goal and what is brief and enduring blocking are matters of judgment, it is quite possible to confuse approach-avoidance conflict with avoidance-approach conflict, unless sufficient information is obtained by observing behavior at different points along time, distance, or cue dimensions.

Perhaps the most interesting difference between approach-avoidance and avoidance-approach conflict is that only in the former case is the individual propelled into the conflict area by his or her own volition and remains trapped there. As a result, in the absence of outside pressure, the person experiences continuous tension. No such stable

equilibrium exists in avoidance-approach conflict.¹ If left to his or her own volition, the individual with an avoidance-approach conflict would simply remain out of the realm of influence of the conflict and would therefore not experience enduring stress. As already noted, this is probably one of the reasons why avoidance-approach conflict has not attracted any attention. Further consideration, however, reveals that avoidance-approach conflict is not as benign as the above analysis suggests. First, consider that the area of conflict to be avoided could be a highly significant one, such as the experience of close relationships with others. The individual with such a conflict could avoid tension only by imposing severe restrictions on his or her realm of experience. It might be argued that the limitations are imposed only by the avoidance gradient, and nothing is added by postulating an avoidance-approach conflict. After all, what function can the approach gradient play if an individual will not voluntarily expose himself or herself to its influence? The answer is that any extrinsic motive or circumstance beyond the individual's control could propel the individual to a point beyond the intersection of the two gradients. In such a circumstance, the person would exhibit behavior that is radically out of character with his or her normal behavior. The person would show an accelerating approach reaction to a goal that had been assiduously avoided up to then.

If the approach motivation involved anti-social behavior, the behavior would appear as an ego-alien breakthrough of a destructive impulse. An example of such a case is the shy, inhibited "model" boy who, suddenly faced with a temptation that he had previously succeeded in avoiding, commits a bizarre crime of passion. On the other hand, if the approach motive were a constructive one, such as the expression of socially acceptable heterosexual feelings and had been avoided because of an excessively broad fear gradient, then an accidental encounter from which withdrawal was difficult could be therapeutic. Once the individual was on the near side of the goal, there would be no further conflict, and the individual would eagerly approach a goal that had previously been avoided. With

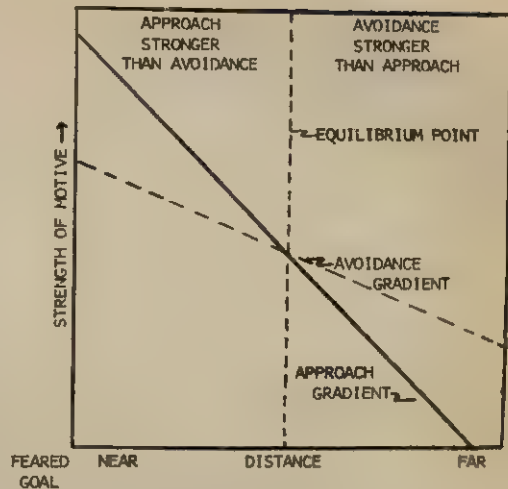


Figure 2. Avoidance-approach conflict. (The individual avoids the goal when at a distance, is momentarily blocked at the point of intersection of the gradients, but, if placed by circumstances beyond his or her control at a closer point will rapidly and intensely approach the goal and exhibit ego-alien behavior.)

repetitive exposure to the same situation, and consequent goal attainment, the conflict would gradually be extinguished.

Given the significant consequences of pathological levels of avoidance-approach conflict, it is important to consider the conditions in real life that could give rise to it. As noted earlier, according to Miller, a critical condition for determining the relative steepness of approach and avoidance gradients is the degree to which response tendencies are influenced by inner relative to external cues. When a high proportion of inner cues is involved in an avoidance gradient, the gradient will be relatively flat and conducive to the development of avoidance-approach conflict. In humans, inner cues frequently consist of thoughts and images. To the extent that an individual is trained to believe that the expression of an impulse in any form is bad, the individual will have a broad avoidance gradient. Now consider that the same individual's approach tendencies are minimally mediated by inner responses, so that if impulses to

¹ This, of course, is also true of approach-approach conflict and of avoidance-avoidance conflict when there are no restraints against leaving the field.

approach are aroused, it is apt to be because of the presence of external cues. Such an individual would have a steeper approach than avoidance gradient, because the avoidance tendencies are internalized to a greater extent than the approach tendencies, and this could even be the case if the approach tendencies were activated by a combination of physiological and external cues, as in the sex drive.

Given sufficiently broad injunctions against the expression of an impulse in any circumstances, there is obviously no need to discriminate between the conditions in which the impulse should or should not be expressed. It will be recalled that poor stimulus discrimination is a major factor in producing broad generalization gradients. Thus, whether one wishes to analyze the comparative steepness of gradients by considering the relative contribution of inner to outer cues, or by considering the role of stimulus discrimination, one arrives at the same conclusion, namely, that strong categorical prohibitions against the expression of an impulse foster broad avoidance gradients and are conducive to the development of an avoidance-approach conflict.

The relative strengths, or heights, of approach and avoidance gradients are obviously critical factors in determining whether an avoidance-approach conflict will occur. If the approach motive is weak enough compared to the avoidance motive, the avoidance motive will, of course, dominate it at all levels, and the individual will simply exhibit a generalized avoidance tendency to all motive-relevant stimuli. The opposite will be the case if the approach motive is sufficiently strong compared to the avoidance motive. Conflict can only exist within a restricted range of relative strengths of the gradients that allows them to intersect. Given such conditions, the stronger the approach and avoidance drives, the more extreme the conflict-related phenomena that will be exhibited, and thus the greater the potential for extreme ego-alien behavior. It follows that the conditions for producing pathological levels of avoidance-approach conflict are those that foster the internalization of intense and broad prohibitions against inherently strong approach motives.

Given the above analysis of avoidance-approach conflict, it follows that overcontrolled persons are more apt to exhibit bizarre ego-alien behavior than undercontrolled persons. This conclusion is supported by studies of criminal violence, which show that the most extreme crimes are committed by individuals whose hostility is characteristically overcontrolled (cf. Blackburn, 1968; Haven, 1973; Megargee, 1966, 1971; White, McAdoo, & Megargee, 1971).

There are some interesting implications for psychotherapy that follow from an analysis of avoidance-approach conflict. For one, if the approach impulses are nondestructive, as in the case of certain sexual feelings, then persuasion, coercion, and confrontation should serve a useful role, as once the individual is brought to the goal side of the point of intersection of the gradients, the conflict will be resolved. This can be contrasted with approach-avoidance conflict, in which the same techniques are contraindicated, as pressure to approach what is feared is apt to induce unmanageable levels of anxiety and therefore cause the patient to withdraw from therapy (cf. Dollard & Miller, 1950). Thus, it is important for the therapist to discriminate between avoidance-approach and approach-avoidance conflict.

It also follows that the therapist must distinguish an avoidance-approach conflict in which the approach motive is socially acceptable from one in which it is destructive. It is obviously unwise to encourage an individual with the latter type of conflict to proceed to the goal side of the intersection point of the gradients or to lower the fear gradient, as either would result in the expression of destructive behavior. It is noteworthy, in this respect, that lowering the fear gradient is the procedure recommended by Dollard and Miller (1950) for treating approach-avoidance conflict.

The method of choice for treating an avoidance-approach conflict when the approach motive is potentially destructive is to increase discrimination of the avoidance dimension so that selective avoidance reactions against extreme and inappropriate expression of the approach motive can be substituted for

a blanket prohibition against its expression in any form.

It is interesting to consider that the differences in therapeutic approaches recommended for approach-avoidance and avoidance-approach conflicts with socially acceptable and unacceptable motives correspond to major emphases in different schools of psychotherapy. For approach-avoidance conflict, the method of choice recommended by Dollard and Miller (1950) is reduction of fear, or lowering of the avoidance motive. This appears to be appropriate when the approach motive involves socially acceptable tendencies. For avoidance-approach conflict, when the approach motive is socially acceptable, the method of choice is to induce the individual to arrive at the goal side of the intersection of the gradients. Although reduction of fear might ultimately work, other procedures, such as encouragement, coercion, and environmental manipulation provide a simpler and more efficient form of treatment. It is noteworthy that such procedures are frowned on by adherents of depth psychology, such as psychoanalysts, and by adherents of accepting forms of treatment, such as client-centered therapists, but are routinely practiced by counseling psychologists, rational-emotive psychologists, and psychodramatists. For both approach-avoidance and avoidance-approach conflicts in which the inhibited drive is destructive, the method of choice is to teach discrimination, both with respect to the stimulus dimension and the response dimension, with the aim of ultimately replacing an all-or-none system of total avoidance or total impulse expression with a system that facilitates modulated control. Schools of therapy that emphasize such an approach among other techniques include psychoanalysis, rational-emotive therapy, and, in some cases, behavior therapy.

References

- Blackburn, R. Personality in relation to extreme aggression in psychiatric offenders. *British Journal of Psychiatry*, 1968, 114, 1301-1302.
- Brown, J. S. Gradients of approach and avoidance responses and their relation to motivation. *Journal of Comparative and Physiological Psychology*, 1948, 41, 450-465.
- Bugelski, B. R., & Woodward, D. P. The effect of distinctive cues on the spatial gradient of avoidance. *Journal of Comparative and Physiological Psychology*, 1951, 44, 450-456.
- Clifford, T. Between-chain and within-chain speed gradients as functions of length. *Learning and Motivation*, 1973, 4, 74-90.
- Desiderato, O., Folds, J. M., & Gockley, J. S. Incubation of anxiety: Effect on temporal generalization. *Psychonomic Science*, 1966, 6, 139-140.
- Dollard, J., & Miller, N. E. *Personality and psychotherapy*. New York: McGraw-Hill, 1950.
- Elder, S. T., Kuehne, T. E., Clarke, N., & Larre, E. E. Approach-avoidance conflict: I. Empirically determined speed-of-locomotion approach gradients. *Psychological Reports*, 1970, 27, 623-628.
- Elder, S. T., Kuehne, T. E., & Moriarty, D. D., Jr. Approach-avoidance conflict: II. Role of olfactory cues. *Psychological Reports*, 1970, 27, 631-638.
- Gewirtz, H. B. Generalization of children's preferences as a function of reinforcement and task similarity. *Journal of Abnormal and Social Psychology*, 1959, 58, 111-118.
- Gjesme, T. Goal distance in time and its effects on the relations between achievement motives and performance. *Journal of Research in Personality*, 1974, 8, 161-171.
- Haven, H. J. Descriptive and developmental characteristics of chronically overcontrolled hostile prisoners. (Doctoral dissertation, Florida State University, 1972.) *Dissertation Abstracts International*, 1973, 33, 5515B.
- Hearst, E. Simultaneous generalization gradients for appetitive and aversive behavior. *Science*, 1960, 132, 1769-1770.
- Hearst, E. Concurrent generalization gradients for food-controlled and shock-controlled behavior. *Journal of the Experimental Analysis of Behavior*, 1962, 5, 19-31.
- Heilizer, F. A review of theory and research on the assumptions of Miller's response competition (conflict) models: response gradients. *Journal of General Psychology*, 1977, 97, 17-71.
- Lewin, K. *A dynamic theory of personality: Selected papers*. New York: McGraw-Hill, 1935.
- Losco, J., & Epstein, S. Relative steepness of approach and avoidance gradients as a function of magnitude and valence of incentive. *Journal of Abnormal Psychology*, 1977, 86, 360-368.
- Maher, B. A. *Principles of psychopathology: An experimental approach*. New York: McGraw-Hill, 1966.
- Megargee, E. I. Undercontrolled and overcontrolled personality types in extreme anti-social aggression. *Psychological Monographs*, 1966, 80(3, Whole No. 611).
- Megargee, E. I. The role of inhibition in the assessment and understanding of violence. In J. E. Singer (Ed.), *The control of aggression and violence: Cognitive and physiological factors*. New York: Academic Press, 1971.
- Miller, N. E. Experimental studies of conflict. In H. McV. Hunt (Ed.), *Personality and the behavior disorders*. New York: Ronald Press, 1944.

- Miller, N. E. Liberalization of basic S-R concepts: Extensions to conflict behavior, motivation and social learning. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 2): *General systematic formulations, learning, and special processes*. New York: McGraw-Hill, 1959.
- Miller, N. E., & Kraeling, D. Displacement: Greater generalization of approach than avoidance in a generalized approach-avoidance conflict. *Journal of Experimental Psychology*, 1952, 43, 217-221.
- Miller, N. E., & Murray, E. J. Displacement and conflict: Learnable drive as a basis for the steeper gradient of avoidance than of approach. *Journal of Experimental Psychology*, 1952, 43, 227-231.
- Murray, E. J., & Berkun, M. M. Displacement as a function of conflict. *Journal of Abnormal and Social Psychology*, 1955, 51, 47-56.
- Rigby, W. K. Approach and avoidance gradients and conflict behavior in a predominantly temporal situation. *Journal of Comparative and Physiological Psychology*, 1954, 47, 83-89.
- Saltz, E., Whitman, R. N., & Paul, C. Performance in the runway as a function of stimulus-differentiation. *American Journal of Psychology*, 1963, 76, 124-127.
- Schroeder, S. R., & Gerjuoy, H. Running-speed goal gradient with massed training. *Psychological Record*, 1965, 15, 425-433.
- Smith, N. W. GSR measures of cigarette smokers' temporal approach and avoidance gradients. *Psychological Record*, 1965, 15, 261-268.
- Smith, N. W. Spatial conflict gradients of cigarette smokers: A methodology. *Journal of General Psychology*, 1969, 80, 287-290.
- Tempone, V. J. Stimulus generalization as a function of mental age. *Child Development*, 1965, 36, 229-235.
- Weiss, R. F. Deprivation and reward magnitude effects on speed throughout the goal gradient. *Journal of Experimental Psychology*, 1960, 60, 384-390.
- White, W. C. Jr., McAdoo, W. G., & Megargee, E. L. Personality factors associated with over- and under-controlled offenders. *FCI Research Reports*, 1971, 3(5), 1-6.

Received July 29, 1977

Effective Ingredients in Psychotherapy: Prediction of Outcome From Process Variables

Beverly Gomes-Schwartz
McLean Hospital

This study was designed to examine the impact of (a) exploration of the psychodynamic roots of patients' conflicts, (b) warmth and friendliness of the therapist-offered relationship, and (c) positiveness of patients' attitudes toward working in therapy on the outcome of brief therapy with 35 college males exhibiting symptoms of depression, anxiety, and social introversion. Analyses of process ratings for audiotaped segments from four sessions throughout the course of therapy revealed that the activities of therapists of differing theoretical orientations and of professional versus untrained, "inherently helpful" therapists could be distinguished. Although patients' attitudes toward the therapist and patient involvement in the therapy process did not differ as a function of the type of therapist, the process dimension that most consistently predicted therapy outcome was patient involvement. Exploratory processes and therapist-offered relationship had a lesser influence on outcome.

Questions about how psychotherapy works—what qualities in the patient, the therapist, and the process of their interaction contribute to the amelioration of the patient's psychic distress—have generated considerable debate among proponents of varying systems of psychotherapy. Some of the issues that have prompted the widest discussion include (a) the relative importance of the patient-therapist relationship, as opposed to specialized techniques of intervention, and (b) the relevance of patients' attitudes toward the therapist and the therapy process.

Relationship Versus Technique

Rogers (1957) asserted that the consistent communication of genuine warmth and empathic understanding by the therapist is sufficient to produce constructive personality change. As long as the therapist is able to offer the patient a warm human relationship, even the most recalcitrant psychotic patients can eventually be reached (Rogers, Gendlin, Kiesler, & Truax, 1967).

In contrast, psychodynamic therapists have emphasized the importance of exploratory techniques—clarification, interpretation, confrontation—for producing the cognitive and emotional insight considered instrumental for change (Bibring, 1954; Glover, 1955; Langs, 1973). Even though a number of analytically oriented therapists have noted the significance of the patient-therapist relationship or therapeutic alliance (Greenson, 1967; Zetzel, 1956), there are fundamental differences between most dynamic therapists and the Rogerians. The analytic therapist is cautioned to maintain the role of an expert healer rather than try to be a friendly or equal partner in an interpersonal relationship. Although the trust and rapport engendered by a good therapeutic relationship may be necessary to facilitate

This article is based on a doctoral dissertation submitted to Vanderbilt University. The research was supported by National Institute of Mental Health Grant 20369 awarded to Hans H. Strupp. The contributions of members of the Vanderbilt Psychotherapy Research Team—Hans H. Strupp, Suzanne W. Hadley, Dianna Hartley, Grady Blackwood, and John Baker—are gratefully acknowledged. Thanks are also due to Joseph M. Schwartz who collaborated in the pilot phases of the research and provided helpful comments, George Silberschatz and Susan Packer who served as raters, and dissertation committee members, Howard B. Roback, Leland E. Thune, John H. Harvey, and Jeffery J. Franks.

Requests for reprints should be sent to Beverly Gomes-Schwartz, Department of Psychology, McLean Hospital, Belmont, Massachusetts 02178.

engagement in an exploratory process, the relationship itself traditionally has not been considered the primary force for change by analytic therapists. (For a divergent view from within the analytic mainstream on the importance of the therapeutic relationship, see Kohut, 1971.)

Empirical analyses of the relative importance of the therapeutic relationship and exploratory techniques in promoting positive change have yielded equivocal results. Although some researchers (e.g., Truax, 1963; Truax & Mitchell, 1971) were initially able to demonstrate that high levels of the relationship variables proposed as necessary and sufficient for therapeutic change by client-centered theorists (i.e., accurate empathy, unconditional positive regard, and congruence) were correlated with therapy outcome, their results have not been replicated in more recent studies (cf. Beutler, Johnson, Neville, & Workman, 1972; Garfield & Bergin, 1971; Mullen & Abeles, 1971; Sloane, Staples, Cristol, Yorkston, & Whipple, 1975). More significantly, serious criticisms of the conceptualization of the scales for measuring "therapist-offered conditions" and the methodology of much of the research have been raised (Chinsky & Rappaport, 1970; Gomes-Schwartz, Hadley, & Strupp, 1978; Gormally & Hill, 1974; Blackwood, Note 1).

Research studying other measures of the therapeutic relationship has also yielded inconsistent results. Adult outpatients in analytically oriented therapy (Feifel & Eells, 1963; Strupp, Fox, & Lessler, 1969) and clients at a college counseling center (Saltzman, Luetgert, Roth, Creaser, & Howard, 1976) who felt that their therapists were warm, understanding, and respectful of them were more likely to be satisfied with their therapy experience and to manifest improvement than were patients who saw their therapists as indifferent, bored, or irritated. In contrast, ratings based on Fiedler's (1950) measure of the "ideal therapeutic relationship" bore a minimal relationship to outcome of analytically oriented group therapy (Parloff, 1961) and no relationship to outcome of individual counseling (Gonyea, 1963; Lesser, 1961).

There has been very little sound research in which either patient or therapist participation in an exploratory, uncovering process was

related to therapy outcome. The value of Malan's (1976) finding that the frequency of interpretations linking early family experiences with the patient-therapist relationship (i.e., transference/parent links) was related to the outcome of brief psychoanalytic therapy is limited by serious methodological deficits, including the use of session notes rather than recordings or transcripts of the interviews and ratings contaminated by raters' knowledge of treatment outcomes. In contrast to Malan's results, among a sample of similar adult outpatients in dynamic therapy, Sloane et al. (1975) found that the frequency of interpretive statements was negatively correlated with improvement on target symptoms. These conflicting results suggest that the value of interpretive techniques may be largely determined by the patient's response to the interpretation. If a patient is generally resistant to self-exploration, even the most perceptive interpretations may be useless.

Unfortunately, the few studies of the relationship between therapy outcome and the patient's engagement in self-exploration (as measured from a client-centered perspective) have yielded equivocal results. Even though Truax and Carkhuff (1967) reported data from several unpublished studies substantiating their hypothesis that greater self-exploration was related to positive outcome, others (Kurtz & Grummon, 1972; Sloane et al., 1975) have found no relationship between self-exploration and outcome.

Patients' Attitudes and Behavior in Therapy

Unlike Rogers (1957), who asserted that virtually any patient could be successfully treated provided the therapist offered sufficient warmth, empathy, and genuineness, many others have suggested that the ways in which patients view their problems and the therapeutic enterprise can influence patients' responses to therapy and, consequently, the benefits that they derive from treatment. Dynamic therapists (Castelnuovo-Tedesco, 1975; Malan, 1976; Sifneos, 1972; Strupp, 1973) have asserted that appropriate therapy candidates must have both the capacity and the motivation to form an intense interpersonal relationship with the therapist and to withstand the stresses of in-

sight-oriented therapy. Although there is some empirical evidence that variables such as ego strength and motivation are related to outcome (see reviews by Gomes-Schwartz et al., 1978; Luborsky, Chandler, Auerbach, Cohen, & Bachrach, 1971; Strupp & Bergin, 1969), the amount of outcome variance that can be predicted from such pretherapy measures has generally been quite small (Auerbach, Luborsky, & Johnson, 1972; Fiske, Cartwright, & Kirtner, 1964).

More striking data on the importance of patient attitudes come from studies of behavior in early therapy sessions. Patients who were involved in the therapy process from the outset of treatment—acknowledging their own responsibility for changing their behavior and actively examining their feelings and experiences—were most likely to improve (Kirtner & Cartwright, 1958; Rice & Wagstaff, 1967; Saltzman et al., 1976). In contrast, patients who viewed their problems as externally imposed or who distanced themselves from the therapy interaction through defensive maneuvers such as intellectualization were unlikely to benefit from therapy (Kirtner & Cartwright, 1958; Rice & Wagstaff, 1967).

Frank (1973) carried the argument for the salience of patient attitudes even farther by suggesting that patients' commitment to change and expectations that therapy can help effect the changes they desire are even more important determinants of therapy outcome than the techniques that a therapist uses. Studies tapping such factors in therapy have yielded some support for Frank's position. Although there is inconsistency in the literature (Wilkins, 1973), a number of investigators have found a significant relationship between patients' expectations that they will benefit from therapy and treatment outcome (e.g., Friedman, 1963; Goldstein & Shipman, 1961; Lipkin, 1954; Martin, Sterne, Moore, & Friedmeyer, 1976). Furthermore, results from several outcome studies suggest that patients (or subjects) who believe that the "placebo" treatments they are receiving are potent therapeutic interventions may experience symptom relief comparable to that of treated patients (Frank, Gliedman, Imber, Stone, & Nash, 1959; Paul, 1966; Smith, 1976).

These results suggest that the content of the

therapist's communications (i.e., interpretation, reflection of feelings) may be less important than the patients' perceptions that they are receiving help. A logical corollary of this idea is that any concerned, helpful listener, with or without professional training as a psychotherapist, can ameliorate psychic distress if he or she offers the patient a warm, supportive relationship and if the patient has confidence in the "therapist's" abilities (Torrey, 1972; Truax & Mitchell, 1971). Empirical data on the effectiveness of nonprofessionals or "inherently helpful" listeners have been conflicting, however. Findings that untrained or minimally trained nonprofessionals were potent therapeutic agents (Carkhuff & Truax, 1965a; Poser, 1966) can be contrasted with findings that interventions by some nonprofessionals actually exacerbated symptoms (Fo & O'Donnell, 1975; Sines, Silver, & Lucero, 1961).

Reflection of Theoretical Differences in the Therapy Process

Just as proponents of client-centered, psychodynamic, and "nonspecific factors" theories of the therapeutic change process differ in their conceptualization of effective therapeutic ingredients, therapists subscribing to varying theoretical systems appear to differ in their practice of psychotherapy. Results from surveys of therapists' usual practices (Rice, Gurman, & Razin, 1974; Sundland & Barker, 1962; Wallach & Strupp, 1964) and from several studies of therapists' behavior in actual or simulated interviews (Strupp, 1955, 1958, 1960) have indicated that analytic therapists emphasized exploratory responses such as questioning and interpretation, whereas client-centered therapists consistently relied on "reflection of feelings." Although data from these studies also indicated that analytic therapists advocated a more formal, professional relationship with their patients than did Rogerian therapists, other investigators found no differences between "nondirective" (or humanistic) and psychodynamic therapists on variables such as warmth, empathy, and genuineness (Fischer, Paveza, Kickert, Hubbard, & Grayston, 1975) or insensitivity,

punitiveness, understanding, and acceptance (Fiedler, 1950).

Comparisons of professional therapists with nonprofessional "helpers" have indicated that psychotherapy training has an important impact on therapeutic techniques. In an analogue of the initial therapy interview, untrained college students emphasized leading responses, particularly direct questions (D'Augelli, Danish, & Brock, 1976). Similarly, Bohn (1965) found that in responding to tape-recorded enactments of a "hostile," "dependent," or "typical" client, relatively experienced graduate student therapists most often used responses categorized as restatement of content or clarification of feelings, whereas naive undergraduate "counselors" relied most heavily on reassurance, persuasion, direct questioning, and forcing the topic. Using a similar set of recordings, Parsons and Parker (1968) found that psychiatric residents were significantly less directive than either senior medical students or college undergraduates.

In contrast to the obtained differences in style of intervention, results from two studies indicated that nonprofessional therapists could not be distinguished from professional therapists on measures of "core conditions." Carkhuff and Truax (1965b) found that with limited training, psychiatric aides were able to offer levels of warmth and empathy comparable to those offered by advanced graduate students and experienced therapists. In initial "therapeutic" interviews, untrained college student volunteers were as warm, genuine, and empathic as experienced psychiatrists and psychiatric residents (Pope, Nudler, VonKorff, & McGee, 1974).

Hypotheses

In summary, there are varying definitions of the effective ingredients in psychotherapy. Client-centered theory emphasizes the curative powers of the good human relationship, whereas psychodynamic theory indicates that a good relationship is not sufficient to induce enduring personality change. Both the psychodynamic perspective and Frank's (1973) concept of nonspecific factors take into account the role of patients' attitudes, whereas client-centered theory carries with it the assumption that all

patients are equally amenable to therapy. In his reformulation of the concepts of the therapeutic influence, Strupp (1973) has suggested that all three elements—the quality of the relationship that the therapist offers, the reconstructive learning experiences that the therapist mediates, and the patient's willingness and capacity to engage in the therapeutic interaction—are determinants of therapeutic change. Given that there have been no attempts to examine the relative contributions of all three factors in a unified research design, the present study was designed to

1. assess the influence of theoretical orientation and psychotherapy training on the therapeutic interactions of analytically oriented, experimental, and untrained, inherently helpful counselors. Experiential therapists and nonprofessional therapists chosen on the basis of their interpersonal skill (identified as alternate therapists in this study) were expected to maintain friendlier, more personal relationships with their patients than analytic therapists. Analytic therapists and their patients were expected to engage in greater exploration of underlying psychodynamics than either experiential or alternate dyads. Alternate therapists were expected to be more directive than professionals of either theoretical orientation.

2. determine the relative impact on the outcome of therapy of (a) engagement in exploratory processes, (b) quality of the therapist-offered relationship, and (c) degree of patient involvement in the therapy interaction. Expectations concerning the relative influence of each dimension may be viewed as tests of competing theories. From the psychodynamic perspective, high levels of exploration (e.g., interpretation, clarification) and high patient involvement (e.g., willingness to communicate, trust in the therapist, recognition of the patient's own responsibility for effecting change) should be the best predictors of outcome. From the Rogerian perspective, high levels of warmth and personal involvement on the part of the therapist should be the most important predictor of change. Finally, from Frank's "nonspecific factors" perspective, both the quality of the therapist-offered relationship and the patient's attitude toward therapy should be the strongest determinants of change.

If the relationship between process and outcome

come is similar across the three groups, this would lend support to the hypothesis that common ingredients account for therapeutic change regardless of the theoretical orientation of the treatment. If, however, the therapists in each treatment group do differ on the process dimensions hypothesized as effective ingredients of psychotherapy and some of these process dimensions are relatively more potent predictors of outcome, analytic, experiential, and alternate therapists should achieve different outcomes.

Method

Subjects

Patients. The patients were 35 unmarried, male college students with elevated scores ($T > 60$) on the Depression (2), Psychasthenia (7), and Social Introversion (0) scales of the Minnesota Multiphasic Personality Inventory¹ (MMPI) who had participated in a psychotherapy outcome study. Patients had been directly referred from a university counseling center or had responded to a letter announcing a special counseling program designed to deal with difficulties in interpersonal relations, anxiety, and shyness. In previous research, similar elevations on the Depression and Psychasthenia scales had proved to be valid indicators of enduring psychological difficulties (Strupp & Bloxom, 1975).

Patients were assigned on a rotational basis to either professional or a nonprofessional therapist. Ten patients were seen by analytic therapists, 10 by experiential therapists, and 15 by alternate therapists. Patients were offered up to 25 sessions of therapy on a once- or twice-a-week basis. Mean durations of therapy for cases seen by analytic, experiential, and alternate therapists were 18.9 ($SD = 6.5$), 16.3 ($SD = 6.9$), and 17.4 ($SD = 7.2$) sessions, respectively.

Therapists. The professional therapists were four male psychiatrists (M experience = 23.5 years) and four male psychologists (M experience = 15.0 years) who were respected clinicians within the Nashville community. The psychiatrists (analytic therapists) all identified psychoanalytic theoreticians as major professional influences. In contrast, the psychologists (experiential therapists) cited the writings of Carl Rogers as having a major impact on their techniques.

The alternate therapists were seven experienced M years since PhD = 17.0) male college professors who had been identified by university administrators, other faculty, and students as teachers who were frequently approached by students for personal counseling. These professors were affiliated with a variety of academic departments including mathematics, English, history, and philosophy. Alternate therapists had been enlisted to participate in a study ostensibly to determine whether inherently helpful people without formal psychotherapy training could aid college students in dealing with their problems in much the same ways that

professional therapists did. They were instructed to behave in the therapy sessions as they usually did when students consulted them for personal advice or counseling. They were specifically advised not to make a special effort to read about psychotherapy.

Instruments

Process scales. To measure the process of psychotherapy, it was necessary to use an instrument that was sufficiently sensitive to capture the quality of the interaction, yet did not require the rating of dimensions so abstract as to preclude reasonable interrater agreement. The Vanderbilt Psychotherapy Process Scale (VPPS; Strupp, Hartley, & Blackwood, Note 2), an 84-item, Likert-type scale, adapted from earlier work by Orlinsky and Howard (1967) to rate the therapy hour from the perspective of a clinical observer, seemed an appropriate instrument.

In an earlier study with a smaller sample of subjects, eight internally consistent subscales derived from the instrument by a priori content analysis successfully discriminated among the three treatment groups—analytic, experiential, and alternate (Gomes-Schwartz & Schwartz, 1978). However, these original scales did not seem to precisely tap the process dimensions that were hypothesized as predictors of therapy outcome. To obtain another perspective on how the individual items might be related, a principal components factor analysis with varimax rotation² was performed on the data from the Gomes-Schwartz and Schwartz (1978) study. Rotated factors with eigenvalues > 1 were defined by items loading $\geq .50$.

From these factors seven scales were derived that tapped dimensions hypothesized as predictors of outcome and that proved to be internally consistent and reliably rated in the present study. (a) Patient Exploration (7 items, coefficient $\alpha = .83$, interrater $r = .88$) gauged the patient's level of self-examination and exploration of feelings and experiences. (b) Therapist Exploration (7 items, $\alpha = .91$, $r = .93$) gauged the degree to which the therapist attempted to examine the psychodynamics underlying the patient's problems. (c) Patient Participation (7 items, $\alpha = .86$, $r = .76$) tapped the degree to which the patient was actively engaged in the therapy interaction (e.g., initiating discussions, not inhibited, etc.). (d) Patient Hostility (6 items, $\alpha = .84$, $r = .82$) measured the level of negativism, hostility, or distrust displayed by the patient. (e) Therapist Warmth and Friendliness (10 items, $\alpha = .83$, $r = .60$) measured the therapist's warmth, caring, and emotional involvement with the patient. (f) Negative Therapist Attitude (3 items, $\alpha = .65$,

¹ Scores on Scales 2, 7, and 0 were not necessarily the only elevations or the highest elevations on the patient's profile. Thus not all of the patients could be categorized as 2-7-0s.

² Although factor analysis is often of questionable value with small samples, it was felt that this technique might generate some meaningful combinations that had been overlooked in the a priori content analysis.

$r = .85$) gauged attitudes that might be expected to threaten or intimidate the patient. (g) Therapist Directiveness (5 items, $\alpha = .88$, $r = .83$) measured directive interventions such as offering advice and modeling behavior. Although this scale was expected to discriminate professional from nonprofessional therapists, a relationship between Therapist Directiveness and treatment outcome was not predicted.

Outcome measures. In previous psychotherapy research (cf. Cartwright, Kirtner & Fiske, 1963; Garfield, Bergin, & Prager, 1971), correlations between outcome indices have generally been low. Therefore, it seemed important to assess treatment effects from several perspectives. Three global and three individualized measures of outcome were selected.

At the conclusion of treatment, both the therapist and an experienced clinical interviewer rated change (6-point Likert-type scales) in (a) severity of the patient's problems; (b) level of the patient's distress; and (c) quality of the patient's functioning in his social, work, and academic roles. Scores on the three items were summed to yield, respectively, therapists' and clinicians' overall ratings of improvement. Global improvement from the patient's perspective was assessed through residual gain scores on an Minnesota Multiphasic Personality Inventory (MMPI) index of malad-

justment (Cooke, 1967). To obtain the MMPI maladjustment score, each of the standard MMPI scales plus Welsh A, Welsh R, and Barron Es were assigned beta weights and summed.

Patients', therapists', and clinicians' ratings of improvement on the target complaints that patients had presented at the outset of therapy were used as more individualized outcome indices. Since patients had been permitted to specify up to three problems that they wished to resolve in therapy, scores were based on the average change across all of the patient's targets.

Procedure

Two advanced clinical psychology graduate students who were not familiar with any of the therapists in the study were selected as raters to reduce the possibility of ratings being biased by knowledge that a therapist was not a professional or by raters' expectations about a therapist based on prior experiences with him.

To provide a broad representative sample of the interaction audiotaped segments from the third interview, the interviews one half and three fourths of the way through treatment and the next-to-last interview were selected for rating. Starting at a randomly deter-

Table 1
Multiple Discriminant Analysis for Process Variables

Process variable	Treatment group			Univariate F^c
	Analytic ^a	Experiential ^a	Alternate ^b	
Patient Exploration				
<i>M</i>	138.70	120.30	93.13	
<i>SD</i>	14.83	16.33	15.93	24.03***
Therapist Exploration				
<i>M</i>	166.10	130.30	90.80	
<i>SD</i>	16.64	16.40	18.80	51.52***
Patient Participation				
<i>M</i>	196.60	192.30	203.87	
<i>SD</i>	24.95	24.90	23.72	<1
Patient Hostility				
<i>M</i>	87.20	100.20	86.00	
<i>SD</i>	17.67	26.35	15.56	1.49
Therapist Warmth and Friendliness				
<i>M</i>	190.40	227.50	224.80	
<i>SD</i>	17.93	29.76	21.65	7.58***
Negative therapist Attitude				
<i>M</i>	43.50	39.60	36.93	
<i>SD</i>	11.01	5.90	13.19	<1
Therapist Directiveness				
<i>M</i>	53.10	88.00	85.47	
<i>SD</i>	12.53	13.72	22.15	11.83***

Note. Wilks $\lambda = .08$; $F(14, 52) = 9.22$, $p < .0001$.

^a $n = 10$.

^b $n = 15$.

^c $df = 2, 32$.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

mined point in each recording (e.g., 11, 19, 13 minutes from the beginning), the succeeding 10-minute segment was recorded onto a master tape such that the segments from all cases and all sessions were presented in random sequence. Independent ratings were made at the conclusion of each 10-minute segment. Since there were no systematic differences in process scores attributable to the time sequence of the segment, all $F_s(3, 96) \leq 1.44$, *ns*, or the interaction of time sequence and treatment group, all $F_s(6, 96) \leq 2.01$, *ns*, overall scores for each process scale were obtained by summing across both the two raters and the four segments.

Results

Differences Among the Treatment Groups

Process variables. To test the hypothesis that therapy interactions differed according to the theoretical orientation or professional-nonprofessional status of the therapist, differences among the three groups—analytic, experiential, and alternate—on the seven process scales were assessed through a multiple discriminant analysis. This analysis, presented in Table 1, yielded a significant overall difference among the groups (Wilks' $\lambda = .08$), $F(14, 52) = 9.22$, $p < .0001$. Examination of differences on individual variables using Dunn's multiple comparison procedure (Dunn, 1961) revealed that the hypothesized relationships among treatment groups were generally obtained.

Both alternate and experiential therapists exhibited greater Therapist Warmth and Friendliness than analytic therapists ($p < .01$). Analytic cases received higher scores than either experiential or alternate cases on Therapist Exploration ($p_s < .05$ and $.01$, respectively) and Patient Exploration ($p_s < .01$ and $.001$, respectively). In addition, the experiential group received higher ratings than the alternates on both Patient Exploration ($p < .01$) and Therapist Exploration ($p < .01$).

The hypothesis that alternate therapists would receive higher ratings than analytic and experiential therapists on Therapist Directiveness received partial support. Alternate therapists were more directive than analytic therapists ($p < .01$). However, experiential therapists did not differ from alternates and were significantly more directive than the analytic group ($p < .01$).

No differences among groups were obtained on ratings of Patient Participation, Patient Hostility, and Negative Therapist Attitude.

Table 2

Multiple Discriminant Analysis for Outcome Variables

Outcome variable ^a	Treatment group		
	Analytic ^b	Experiential ^b	Alternate ^c
Overall ratings			
Clinicians			
<i>M</i>	6.20	5.15	5.40
<i>SD</i>	3.51	3.84	2.84
Therapists			
<i>M</i>	6.10	5.30	4.93
<i>SD</i>	3.77	4.65	2.23
MMPI mal-adjustment			
<i>M</i>	50.03	50.82	49.43
<i>SD</i>	11.41	8.18	10.05
Target complaints			
Patients			
<i>M</i>	2.56	1.92	2.19
<i>SD</i>	1.27	1.65	1.68
Therapists			
<i>M</i>	1.76	1.81	2.36
<i>SD</i>	1.28	1.64	1.12
Clinicians			
<i>M</i>	1.73	1.69	1.97
<i>SD</i>	1.25	1.79	1.90

Note. Wilks $\lambda = .69$; $F(12, 54) = .92$, *ns*. All univariate $F_s(2, 32) < 1$.

^a For all variables except Minnesota Multiphasic Personality Inventory (MMPI) maladjustment, higher scores indicate greater change.

^b $n = 10$.

^c $n = 15$.

Outcome variables. A discriminant function analysis comparing the treatment groups on the six outcome criteria yielded neither a significant overall difference (Wilks $\lambda = .69$), $F(12, 54) = .92$, *ns*, nor differences on individual outcome criteria (see Table 2).

Prediction of Outcome from Process Dimensions

The relationship between psychotherapy process and outcome was examined in several ways to determine (a) which process dimensions (i.e., exploratory processes, patient involvement, therapist-offered relationship) were the best predictors of outcome, (b) whether the relationships between process dimensions and outcome variables were primarily due to the contribution of individual dimensions or

Table 3
Multiple Regression Predicting Outcome from Process Dimensions

Outcome variable	Exploratory processes			Patient involvement			Therapist-offered relationship		
	<i>R</i>	<i>R</i> ²	<i>F</i>	<i>R</i>	<i>R</i> ²	<i>F</i>	<i>R</i>	<i>R</i> ²	<i>F</i>
Overall ratings									
Clinicians	.26	.04	1.17	.54	.27	6.73**	.28	.05	1.32
Therapists	.45	.18	4.10*	.56	.30	7.46**	.33	.08	1.97
MMPI maladjustment	.13	-.01	<1	.29	.05	1.44	.30	.06	1.62
Target complaints									
Patients	.14	-.01	<1	.38	.12	2.77	.10	-.02	<1
Therapists	.39	.12	2.83	.63	.38	10.80***	.51	.24	5.56**
Clinicians	.10	-.02	<1	.38	.12	2.70	.27	.05	1.31

Note. MMPI = Minnesota Multiphasic Personality Inventory. *d*fs for *F* tests = 2, 32. Reported *R*² is the unbiased estimate of population *R*².

* $p < .05$.

** $p < .01$.

*** $p < .001$.

to the interactions among the dimensions, and (c) whether the relationship between process and outcome was consistent across treatment groups.

Since each of the process dimensions hypothesized as a predictor of outcome was tapped by two process scales (i.e., exploratory processes = Patient Exploration and Therapist Exploration; patient involvement = Patient Participation and Patient Hostility; therapist-offered relationship = Therapist Warmth and Friendliness and Negative Therapist Attitude), multiple regression analyses predicting each of the outcome variables from each pair of process scales were performed. Results from these analyses, presented in Table 3, indicated that patient involvement was consistently the best predictor of outcome. Multiple correlations of patient involvement with clinicians' and therapists' overall improvement ratings and therapists' ratings of improvement in target complaints were significant. In addition, multiple correlations of involvement with improvement in patients' and clinicians' target complaints approached significance ($p < .10$). In contrast, exploratory processes predicted only therapists' overall improvement ratings ($p < .05$), and therapist-offered relationship predicted only improvement in therapists' target complaints ($p < .01$).

Not only did patient involvement bear a significant relationship with more outcome variables than the process dimensions, in each

case in which another process dimension was significantly correlated with outcome, but patient involvement accounted for more of the variance in outcome ratings (i.e., 30% vs. 18% for exploratory processes and 38% vs. 24% for therapist-offered relationship).

Since the three process dimensions cannot be regarded as independent (e.g., the patient's attitude may be influenced by the therapist's behavior), it was necessary to determine whether the significant multiple correlations represented the interactive effects of more than one process dimension or were primarily determined by a single dimension. Thus, partial correlations between each process dimension and each outcome variable with the effects of the remaining two process dimensions partialled out were computed (see Table 4). These analyses revealed that patient involvement exclusive of the influences of both exploratory processes and therapist-offered relationship showed a consistent relationship with outcome. Four of the six partial correlations between involvement and outcome variables were significant. In contrast, none of the partial correlations between exploratory processes or therapist-offered relationship and outcome approached significance.

To determine whether the observed relationships between process and outcome were similar for the three treatment groups (analytic, experiential, and alternate), it was necessary to test whether the multiple regression slopes

Table 4
Partial Correlations Between Process Dimensions and Outcome Variables

Outcome variable	Exploratory processes	Patient involvement	Therapist-offered relationship
Overall ratings			
Clinicians	.08	.49**	-.05
Therapists	.20	.50**	-.14
MMPI maladjustment	-.03	.29	-.12
Target complaints			
Patients	.14	.38*	-.14
Therapists	.16	.54**	.18
Clinicians	-.02	.25	.08

Note. N for all correlations = 35, df = 31. Scores for process dimensions are weighted combinations of process scale scores based upon beta values from multiple regression analyses. MMPI = Minnesota Multiphasic Personality Inventory.

* $p < .05$.

** $p < .01$.

depicting the relation of the three process dimensions to each outcome variable differed across treatment groups. F tests for homogeneity of regression yielded no significant effects, all F s(6, 23) \leq 1.59, ns , indicating that the relationship between process and outcome was consistent regardless of the theoretical orientation or professional status of the therapist.

Discussion

The primary findings in this study were as follows: (a) Theoretical orientation and professional-nonprofessional status of the therapist had an impact on the process but not the outcome of psychotherapy; and (b) therapy outcome was most consistently predicted by the patient's willingness and ability to become actively involved in the therapy interaction—a dimension of therapy process that did *not* distinguish among the three treatment groups.

As predicted, professionally trained therapists, particularly those with an analytic orientation, and their patients invested more effort in uncovering the psychodynamic roots of the patient's problems than did the dyads led by nonprofessional therapists. Therapists with experiential training and the untrained counselors offered warmer, more personal relationships with their patients than did therapists who assumed an analytic stance. These findings are largely consistent with the psy-

chotherapeutic theories to which the therapists subscribed. The analyst has traditionally been taught to remain aloof—an expert healer rather than a warm friend (Langs, 1973). In contrast, warmth, empathy, and genuineness are considered to be the fundamental tools of the client-centered therapist (Rogers, 1957). Perhaps the emergence of "self" theory in psychoanalysis with its increased focus on the importance of the therapist's capacity to respond empathically (see, for example, Kohut, 1971) may eventually challenge traditional notions concerning the appropriate analytic posture and blur some of the distinctions between client-centered and dynamic therapists' self-presentations in therapy. However, as the behavior of the analytic therapists in this study may illustrate, the therapist who has assumed the traditional passive, non-demonstrative role for many years may find it difficult to relinquish.

Even though the therapists for each of the three treatment groups behaved quite differently, these differences did not seem to influence patients' attitudes toward therapy or the therapist. Patients were as likely to become involved in the therapy process regardless of whether they saw analytic, experiential, or alternate therapists. The fact that the patient's willingness to ally himself with the therapist and work at changing was not influenced by the theoretical orientation and professional status of his therapist may be of particular importance for understanding why there were

no differences among the groups on treatment outcome. These patient characteristics that did *not* distinguish the three treatment groups consistently emerged as the best predictors of therapy outcome. Patients who were not hostile or mistrustful and who actively contributed to the therapy interaction achieved greater changes than those who were withdrawn, defensive, or otherwise unwilling to engage in the therapy process.

How can the findings from this study best be accommodated with theories of therapeutic influence? The results are clearly congruent with Frank's (1973) theory of "nonspecific factors" as the determinants of outcome. Given mildly to moderately disturbed patients in short-term therapy (an average of 17.4 sessions), untrained professor/therapists generally affected as much improvement as experienced psychologists and psychiatrists. Further, the variables that best predicted change were not related to therapeutic techniques but to the positiveness of the patient's attitude toward his therapist and his commitment to work at changing.

The present findings are also consonant with the ideas of specialists in brief psychoanalytic therapy that those patients who have the willingness and the adaptive resources to work with the therapist to resolve their problems are most likely to profit from short-term therapy (Castelnuovo-Tedesco, 1975; Malan, 1976; Sifneos, 1972). Although there is little evidence in this study that the use of exploratory technique also strongly influenced outcome as analytic theory would suggest, it is possible that technique variables might have been more salient if treatment were of longer duration (e.g., the 40-session limit proposed by Malan, 1976, or the maximum of 1 year suggested by Sifneos, 1972) or if only patients "appropriate" for brief dynamic therapy (i.e., those with sufficient motivation and ego resources) were selected for treatment. If difficulties in maintaining self-esteem, which a considerable portion of the patients in this study exhibited, can be viewed as indicators of narcissistic pathology, Kohut's (1971) proposal that patients with severe pathology in the area of the "self" benefit less from dynamic interpretation than from the empathic attention of the therapist may also aid in explaining

why proficiency in dynamic exploration bore so little relation to outcome in this sample.

If the patient's capacity and willingness to participate in the therapy interaction are among the most important determinants of improvement in short-term therapy, one of the aims of future research should be to determine if and how positive involvement in therapy could be increased. It is essential to determine the extent to which the patient's negativism or inability to collaborate reflects (a) relatively stable personality characteristics or (b) response to psychonoxious attitudes and behaviors on the part of the therapist.

If as some therapists (Castelnuovo-Tedesco, 1975; Sifneos, 1972) and researchers (Kirtner & Cartwright, 1958; Rice & Wagstaff, 1967) have suggested, the patient's ability to actively work toward resolving problems is a reflection of an enduring character structure or life view, researchers may be able to detect evidence of a generalized sense of hostility or passive indifference in the patient's interactions with others (e.g., an intake interviewer). Finding that the degree to which patients become positively involved in the therapy process was influenced by long-standing personality characteristics would have important implications for therapy practice.

One option for maximizing the effectiveness of psychotherapy would be to select only those patients who evidence a capacity to actively participate in a therapeutic interaction. Rather than offering psychotherapy, particularly uncovering therapy, to all patients who present themselves at a clinic or community mental health center, clinicians might consider alternative interventions for applicants who did not appear to have the capacity to ally themselves with a therapist or to assume a great deal of the responsibility for changing themselves.

Another approach might be to alter some of the behaviors that prevent patients from succeeding in psychotherapy. If patients' abilities to become involved in the therapy process were as much a product of inappropriate expectations about the psychotherapy enterprise as a lack of willingness to take responsibility for their own behavior, role-induction procedures (cf. Hoehn-Saric et al., 1964; Strupp & Bloxom, 1973) might be useful to demonstrate to pro-

spective patients what types of behaviors would be expected of them in therapy.

Even if enduring attitudes were a major determinant of a patient's capacity to become involved in therapy, there are very likely some therapist behaviors that can prompt or exacerbate negativism, defensiveness, or apathy. Determining the therapist behaviors, attitudes, or character traits that interact with patient attitudes to influence involvement may have important implications for the selection and training of therapists and for the "matching" of optimal patient-therapist pairs.

In conclusion, the results of this study have illustrated that meaningful measures of psychotherapy process are possible, and that outcome can be predicted from the process of the therapy interaction. However, additional questions about the role of the patient and the therapist in determining the course of therapy have been raised. It is reasonable to hope that some of these questions can be answered through continued research efforts.

Reference Notes

1. Blackwood, G. L., Jr. *Accurate empathy: Critique of a construct*. Unpublished manuscript, Vanderbilt University, 1975.
2. Strupp, H. H., Hartley, D., & Blackwood, G. L., Jr. *Vanderbilt Psychotherapy Process Scale*. Unpublished manuscript, Vanderbilt University, 1974.

References

- Auerbach, A., Luborsky, L., & Johnson, M. Clinicians' predictions of psychotherapy outcome: A trial of a prognostic index. *American Journal of Psychiatry*, 1972, 128, 830-835.
- Beutler, L. E., Johnson, P. T., Neville, C. W., & Workman, S. N. "Accurate empathy" and the A-B dichotomy. *Journal of Consulting and Clinical Psychology*, 1972, 38, 372-375.
- Bibring, E. Psychoanalysis and the dynamic psychotherapies. *Journal of American Psychoanalytic Association*, 1954, 2, 745-770.
- Bohn, M. J., Jr. Counselor behavior as a function of counselor dominance, counselor experience, and client type. *Journal of Counseling Psychology*, 1965, 12, 346-352.
- Carkhuff, R. R., & Truax, C. B. Lay mental health counseling: The effects of lay group counseling. *Journal of Consulting Psychology*, 1965, 29, 426-431. (a)
- Carkhuff, R. R., & Truax, C. B. Training in counseling and psychotherapy: An evaluation of an integrated and experiential approach. *Journal of Consulting Psychology*, 1965, 29, 333-336. (b)
- Cartwright, D. S., Kirtner, W. L., & Fiske, D. W. Method factors in changes associated with psychotherapy. *Journal of Abnormal and Social Psychology*, 1963, 66, 164-175.
- Castellnuovo-Tedesco, P. Brief psychotherapy. In S. Arieti (Ed.), *American handbook of psychiatry* (2nd ed., Vol. 5). New York: Basic Books, 1975.
- Chinsky, J. M., & Rappaport, J. Brief critique of the meaning and reliability of "accurate empathy" ratings. *Psychological Bulletin*, 1970, 73, 379-380.
- Cooke, J. K. MMPI in actuarial diagnosis of psychological disturbance among college males. *Journal of Counseling Psychology*, 1967, 14, 474-477.
- D'Augelli, A. R., Danish, S. J., & Brock, G. W. Untrained paraprofessionals' verbal helping behavior: Description and implications for training. *American Journal of Community Psychology*, 1976, 4, 275-282.
- Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, 1961, 56, 52-64.
- Feifel, H., & Eells, J. Patients and therapists assess the same psychotherapy. *Journal of Consulting Psychology*, 1963, 27, 310-318.
- Fiedler, F. E. The concept of an ideal therapeutic relationship. *Journal of Consulting Psychology*, 1950, 14, 239-245.
- Fischer, J., Paveza, G. J., Kickert, N. S., Hubbard, L. J., & Grayston, S. B. The relationship between theoretical orientation and therapists' empathy, warmth, and genuineness. *Journal of Counseling Psychology*, 1975, 22, 399-403.
- Fiske, D. W., Cartwright, D. S., & Kirtner, W. L. Are psychotherapeutic changes predictable? *Journal of Abnormal and Social Psychology*, 1964, 69, 418-426.
- Fo, W. S., & O'Donnell, C. R. The buddy system: Effect of community intervention on delinquent offenses. *Behavior Therapy*, 1975, 6, 522-524.
- Frank, J. D. *Persuasion and healing*. Baltimore, Md.: Johns Hopkins University Press, 1973.
- Frank, J. D., Gliedman, L. H., Imber, S. D., Stone, A. R., & Nash, E. H. Patients' expectations and relearning as factors determining improvement in psychotherapy. *American Journal of Psychiatry*, 1959, 115, 961-968.
- Friedman, H. J. Patient expectancy and symptom reduction. *Archives of General Psychiatry*, 1963, 8, 61-67.
- Garfield, S. L., & Bergin, A. E. Therapeutic conditions and outcome. *Journal of Abnormal Psychology*, 1971, 77, 108-114.
- Garfield, S. L., Bergin, A. E., & Prager, R. A. Evaluation of outcome in psychotherapy. *Journal of Consulting and Clinical Psychology*, 1971, 37, 307-313.
- Glover, E. *The techniques of psychoanalysis*. New York: International Universities Press, 1955.
- Goldstein, A. P., & Shipman, W. G. Patient expectancies, symptom reduction, and aspects of the initial psychotherapeutic interview. *Journal of Clinical Psychology*, 1961, 17, 129-133.
- Gomes-Schwartz, B., Hadley, S. W., & Strupp, H. H. Psychotherapy and behavior therapy. In *Annual review of psychology*. Palo Alto, Calif.: Annual Reviews, 1978.

- Gomes-Schwartz, B., & Schwartz, J. M. Psychotherapy process variables distinguishing the "inherently helpful" person from the professional psychotherapist. *Journal of Consulting and Clinical Psychology*, 1978, 46, 196-197.
- Gonyea, G. C. The "ideal therapeutic relationship" and counseling outcome. *Journal of Clinical Psychology*, 1963, 19, 481-487.
- Gormally, J., & Hill, C. E. Guidelines for research on Carkhuff's training model. *Journal of Counseling Psychology*, 1974, 21, 539-547.
- Greenson, R. *The technique and practice of psychoanalysis*. New York: International Universities Press, 1967.
- Hoehn-Saric, R., et al. Systematic preparation of patients for psychotherapy: Effects on therapy behavior and outcome. *Journal of Psychiatric Research*, 1964, 2, 267-281.
- Kirtner, W. L., & Cartwright, D. S. Success and failure in client-centered therapy as a function of initial in-therapy behavior. *Journal of Consulting Psychology*, 1958, 22, 329-333.
- Kohut, H. *The analysis of the self*. New York: International Universities Press, 1971.
- Kurtz, R. R., & Grummon, D. L. Different approaches to the measurement of therapist empathy and their relationship to therapy outcome. *Journal of Consulting and Clinical Psychology*, 1972, 39, 106-115.
- Langs, R. *The technique of psychoanalytic psychotherapy*. New York: Aronson, 1973.
- Lesser, W. M. The relationship between counseling proper and empathic understanding. *Journal of Counseling Psychology*, 1961, 8, 330-336.
- Lipkin, S. Client feelings and attitudes in relation to the outcome of client-centered therapy. *Psychological Monographs*, 1954, 68(1, Whole No. 372).
- Luborsky, L., Chandler, M., Auerbach, A. H., Cohen, J., & Bachrach, H. M. Factors influencing the outcome of psychotherapy: A review of quantitative research. *Psychological Bulletin*, 1971, 75, 145-185.
- Malan, D. H. *Towards the validation of dynamic psychotherapy: A replication*. New York: Plenum Medical Books, 1976.
- Martin, P. J., Sterne, A. L., Moore, J. E., & Friedmeyer, M. H. Patients' and therapists' expectancies and treatment outcome: An elusive relationship re-examined. *Research Communications in Psychology, Psychiatry, and Behavior*, 1976, 1, 301-314.
- Mullen, J., & Abeles, N. Relationship of liking, empathy, and therapist's experience to outcome of therapy. *Journal of Counseling Psychology*, 1971, 18, 39-43.
- Orlinsky, D. E., & Howard, K. I. The good therapy hour. *Archives of General Psychiatry*, 1967, 16, 621-632.
- Parloff, M. B. Therapist-patient relationship and outcome of psychotherapy. *Journal of Consulting Psychology*, 1961, 25, 29-38.
- Parsons, L. B., & Parker, G. V. Personal attitudes, clinical appraisals, and verbal behavior of trained and untrained therapists. *Journal of Consulting and Clinical Psychology*, 1968, 32, 64-71.
- Paul, G. L. *Insight versus desensitization in psychotherapy: An experiment in anxiety reduction*. Stanford, Calif.: Stanford University Press, 1966.
- Pope, B., Nudler, S., VonKorff, M. R., & McGee, J. P. The experienced professional interviewer versus the complete novice. *Journal of Consulting and Clinical Psychology*, 1974, 42, 680-690.
- Poser, E. G. The effects of therapists' training on group therapeutic outcome. *Journal of Counseling Psychology*, 1966, 30, 283-289.
- Rice, D. G., Gurman, A. S., & Razin, A. M. Therapist sex, style, and theoretical orientation. *Journal of Nervous and Mental Disease*, 1974, 159, 413-421.
- Rice, L. N., & Wagstaff, A. K. Client voice quality and expressive style as indices of productive psychotherapy. *Journal of Consulting Psychology*, 1967, 31, 557-563.
- Rogers, C. R. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 1957, 21, 95-103.
- Rogers, C. R., Gendlin, E. T., Kiesler, D., & Truax, C. B. *The psychotherapeutic relationship and its impact: A study of psychotherapy with schizophrenics*. Madison: University of Wisconsin Press, 1967.
- Saltzman, C., Luetgert, M. J., Roth, C. H., Creaser, J., & Howard, L. Formation of a therapeutic relationship: Experiences during the initial phase of psychotherapy as predictors of treatment duration and outcome. *Journal of Consulting and Clinical Psychology*, 1976, 44, 546-555.
- Sifneos, P. *Short-term psychotherapy and emotional crisis*. Cambridge, Mass.: Harvard University Press, 1972.
- Sines, L. K., Silver, R. J., & Lucero, R. J. The effect of therapeutic intervention by untrained "therapists." *Journal of Clinical Psychology*, 1961, 17, 394-396.
- Sloane, R. B., Staples, F. R., Cristol, A. H., Yorkston, N. J., & Whipple, K. *Psychotherapy versus behavior therapy*. Cambridge, Mass.: Harvard University Press, 1975.
- Smith, J. C. Psychotherapeutic effects of transcendental meditation with controls for expectation of relief and daily sitting. *Journal of Consulting and Clinical Psychology*, 1976, 44, 630-637.
- Strupp, H. H. An objective comparison of Rogerian and psychoanalytic techniques. *Journal of Consulting Psychology*, 1955, 19, 1-17.
- Strupp, H. H. The performance of psychoanalytic and client-centered therapists in an initial interview. *Journal of Consulting Psychology*, 1958, 22, 265-274.
- Strupp, H. H. *Psychotherapists in action*. New York: Grune & Stratton, 1960.
- Strupp, H. H. Toward a reformulation of the psychotherapeutic influence. *International Journal of Psychiatry*, 1973, 11, 263-327.
- Strupp, H. H., & Bergin, A. E. Some empirical and conceptual bases for coordinated research in psychotherapy. *International Journal of Psychiatry*, 1969, 7, 18-90.
- Strupp, H. H., & Bloxom, A. L. Preparing lower-class patients for group psychotherapy: Development and evaluation of a role-induction film. *Journal of Consulting and Clinical Psychology*, 1973, 41, 373-384.
- Strupp, H. H., & Bloxom, A. L. An approach to defining a patient population in psychotherapy.

- research. *Journal of Counseling Psychology*, 1975, 22, 231-237.
- Strupp, H. H., Fox, R. E., & Lessler, K. *Patients view their psychotherapy*. Baltimore, Md.: Johns Hopkins University Press, 1969.
- Sundland, D., & Barker, E. N. The orientations of psychotherapists. *Journal of Consulting Psychology*, 1962, 26, 201-212.
- Torrey, E. F. *The mind game*. New York: Bantam Books, 1972.
- Truax, C. B. Effective ingredients in psychotherapy. *Journal of Counseling Psychology*, 1963, 10, 256-263.
- Truax, C. B., & Carkhuff, R. R. *Toward effective counseling and psychotherapy: Training and practice*. Chicago: Aldine, 1967.
- Truax, C. B., & Mitchell, K. M. Research on certain therapist interpersonal skills in relation to process and outcome. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change*. New York: Wiley, 1971.
- Wallach, M. S., & Strupp, H. H. Dimensions of psychotherapists' activities. *Journal of Consulting Psychology*, 1964, 28, 120-125.
- Wilkins, W. Expectancy of therapeutic gain: An empirical and conceptual critique. *Journal of Consulting and Clinical Psychology*, 1973, 40, 69-77.
- Zetzel, E. R. Current concepts of transference. *International Journal of Psychoanalysis*, 1956, 37, 369-376.

Received August 8, 1977 ■

Coping and the Self-Control of Chronic Tension Headache

Kenneth A. Holroyd and Frank Andrasik
Ohio University

Thirty-nine community residents with chronic tension headache were assigned to one of two self-control treatment groups, a headache discussion group, or a symptom-monitoring control group. Participants in the two self-control treatment groups and in the headache discussion group were provided similar rationales for treatment and were taught to monitor their cognitive responses to stress-eliciting situations. Participants in the two self-control treatment groups were also taught either cognitive or both cognitive and relaxation coping skills for controlling tension headache. Participants in the headache discussion group were not provided with specific skills for controlling their headaches but were led in a discussion of the historical roots of their symptoms. Both the self-control treatments and the headache discussion procedure produced substantial reductions in headache that were maintained at a 6-week follow-up. The symptom-monitoring control group showed no change in headache symptoms. These findings provide additional evidence of the effectiveness of cognitively oriented therapeutic procedures for the treatment of tension headache but raise questions concerning the active ingredients of these treatments.

Headache may be the most commonly reported bodily complaint (Wolff, 1963). Survey data indicate that between 50% and 70% of adults experience headaches, 40% of which are tension headaches (Kashiwagi, McClure, & Wetzel, 1972). Of the 15 classes of headache identified by the Ad Hoc Committee on Classification of Headache (1962) of the American Medical Association, tension headache, also commonly termed muscle contraction, psychogenic, or nervous headache, is the most frequently occurring. Tension headache is typically characterized by persistent sensations of bandlike pain or tightness located bilaterally in the occipital and/or forehead regions. It is gradual in onset and may last for hours, weeks, or even months.

The exact etiology of tension headache

remains unclear (Bakal, 1975). However, there is a general consensus that tension headache (a) is an individual response to psychological stress (Ad Hoc Committee on the Classification of Headache, 1962; Wolff, 1963) and (b) may result from the sustained contraction of skeletal muscles about the face, scalp, neck, and shoulders (Bakal, 1975; Martin, 1972).

Behavioral approaches to the treatment of tension headache have focused on modifying the muscle contraction responses presumed to contribute to tension headache (Budzynski, Stoyva, Adler, & Mullaney, 1973; Cox, Freudlich, & Meyer, 1975; Haynes, Griffin, Mooney, & Parise, 1975; Hutchings & Reinking, 1976). However, a somewhat different approach was taken by Holroyd, Andrasik, and Westbrook (1977), who found that a self-control treatment that focused on modifying cognitive responses to stress-eliciting situations was more effective in reducing tension headaches than biofeedback-assisted relaxation training when these treatments were accompanied by counterdemand instructions. The self-control treatment used in Holroyd et al. contained cognitively oriented thera-

Appreciation is expressed to Dale Mattmiller, Henrietta Russell, and Teresa Abell who assisted with the carrying out of this experiment. This study was supported by Grant 1-R03-MH28939-1 from the National Institute of Mental Health to the first author.

Requests for reprints should be sent to Kenneth A. Holroyd, Department of Psychology, Ohio University, Athens, Ohio 45701.

peutic procedures (Beck, 1976; Goldfried, Decentecio, & Weinberg, 1974; Meichenbaum, 1977) to teach individuals to identify their reactions to stress and to use effective cognitive coping skills. Similar treatment procedures have been found effective not only for the treatment of clinical problems such as specific anxieties (Di Loreto, 1971; Goldfried, Linehan, & Smith, 1978; Holroyd, 1976; Meichenbaum, 1972; Meichenbaum, Gilmore, & Fedoravicious, 1971; Kanter & Goldfried, Note 1), depression (Rush, Beck, Kovacs, & Hollon, 1977), stuttering (Moleski & Tosi, 1976), and unassertive behavior (Thorpe, 1975; Wolfe & Fodor, 1977; Linehan & Goldfried, Note 2) but also for providing individuals with skills for coping with laboratory (Meichenbaum, Turk, & Burnstein, 1975) and real-life (Langer, Janis, & Wolfer, 1975) stressors.

In light of growing evidence of the effectiveness of such cognitive self-control treatments, it becomes important to determine to what extent the various components of these treatments contribute to the effectiveness of the treatments. Although such complex social influence procedures are difficult to dissect, recent analyses (Beck, 1976; Goldfried, 1977; Holroyd et al., 1977; Meichenbaum, 1977) suggest that these treatment procedures at least influence clients to (a) attribute the source of their symptoms to relatively specific cognitive aberrations rather than to external stimuli or complex inner dispositions; (b) identify or self-monitor a cognitive component of their distress; and (c) engage in specific cognitive coping strategies (e.g. reappraisal, self-instruction, imagery, etc.). The present study attempted to determine to what extent the specific coping strategies that are taught during treatment contribute to treatment outcome.

Chronic headache sufferers were assigned to one of two self-control treatment groups, a headache discussion group, or a symptom-monitoring control group. Participants in the two self-control treatment groups and in the headache discussion group were provided similar rationales for treatment designed to influence them to adopt similar cognitively oriented explanations for their symptoms, and were taught to monitor their cognitive

responses to stress-eliciting situations in a similar manner. Participants in the two self-control treatment groups were also taught either cognitive coping skills or both cognitive and relaxation skills for managing stress and controlling tension headache. Participants in the headache discussion group were provided no specific skills for managing stress or controlling headaches. However, since previous work has indicated that a treatment procedure consisting solely of the treatment rationale and instruction in the monitoring of cognitive responses to stress lacks credibility, an alternate therapeutic task involving a discussion of the historical sources of headache symptoms was provided for this group. Symptom-monitoring control group members recorded their headaches but did not receive treatment until the study had been completed.

Method

Subjects

Articles in local newspapers, appearances on radio "talk shows," and wall posters were used in addition to newspaper, radio, and television advertisements to circulate announcements of a program teaching methods for the self-control of tension headache to communities within a 50-mile radius of Athens, Ohio. Telephone and small group screening procedures, as well as evaluations by participants' physicians, were used to identify those individuals exhibiting clear-cut symptoms of tension headache occurring consistently at least three times per week from 123 initial respondents. Participants included 35 females and 4 males with a mean age of 35.2 years and a mean duration of headache symptoms of 10.1 years.

Procedure

Following initial telephone screening, potential participants were seen in small groups by one of the authors, who obtained informed consent and a required \$5 deposit, administered pretreatment measures, and arranged to obtain diagnostic medical information. A list of 19 characteristics of tension headache was also used to eliminate volunteers suffering from other types of headache or reporting mixed headache symptoms (Wolff, 1963). Tension headache sufferers who reported consistently experiencing a minimum of three tension headaches per week were assigned by a within-sample matching technique (Goldstein, Heller, & Sechrest, 1966) to one of the two self-control treatments, the headache discussion group, or to the symptom-monitoring control group.

Measures

Symptoms. For at least 2 weeks prior to receiving treatment, continuing through a 2-week post-treatment assessment, and for 2 weeks at a 6-week follow-up evaluation, participants maintained a headache data card on which they rated the occurrence and intensity of headaches on an 11-point scale (0 = no headache, 10 = incapacitating headache) every hour from 10:00 a.m. through 10:00 p.m. (Haynes et al., 1975). An index of overall headache activity was calculated by summing headache ratings each day with weekly averages.¹ In addition, participants' ratings of the frequency of occurrence of 18 common psychosomatic complaints were obtained on the psychosomatic checklist (Cox et al., 1975) at pretreatment, posttreatment, and follow-up evaluations.

Frontalis electromyographic (EMG) activity. Frontalis EMG was assessed during resting periods at pretreatment evaluation and, for participants receiving treatment, following each treatment session. While participants rested in a heavily padded chair, in a different location from where they received treatment, frontalis muscle activity was monitored from forehead disk electrode placements (Budzynski et al., 1973) directed to a Cyborg BL933 electromyograph. Three 1-minute integrated EMG measurements were sampled and averaged to provide an index of frontalis muscle activity.

Additional measures. To assess their perceptions of the credibility of the treatment that they received, the participants evaluated the probability of their recommending the treatment to a friend suffering from tension headache and how important they felt it was that their treatment be made available to other headache sufferers (on 5-point scales) following the first treatment session, at posttreatment, and at follow-up evaluations. Participants also rated the warmth, empathy, and skill of their therapist at the posttreatment and follow-up evaluations. Since it was suspected that the headache discussion procedure might increase participants' levels of self-esteem without necessarily reducing specific symptoms, self-esteem was assessed by the Miskimins Self-Goal Other Discrepancy Scale (Ryan, Krall, & Hodges, 1976) at all three assessments. No attempts were made to influence participants' medication intake. However, medication intake was recorded on a daily basis on the headache data cards.

Treatment

Treatment procedures were administered during five weekly 1½-hour group meetings. We each conducted one group of each type so that therapists and groups were completely crossed. Although we both have had previous experience in treating chronic tension headaches with self-control procedures in an individual format (e.g., Holroyd et al., 1977), the first author had somewhat more group therapy experience (5 years) than the second author, who had conducted one group previous to this study. Counter-demand instructions emphasizing that no improve-

ment could be expected until the completion of treatment (Steinmark & Borkovec, 1974) were administered at the end of the first treatment session.

Cognitive self-control ($n=10$). This treatment focused on altering maladaptive cognitive responses that were assumed to mediate the occurrence of tension headache. The treatment format closely followed that used by Holroyd et al. (1977), with appropriate modifications for the group setting. Specific procedures were adapted from cognitively oriented therapy procedures (Beck, 1976; Goldfried et al., 1974; Meichenbaum, 1977) and were designed to maximize the occurrence of causal reattribution and the development of self-monitoring and cognitive coping skills described below.

The rationale for treatment emphasized that disturbing emotional and behavioral responses are a direct function of specifiable maladaptive cognitions. It was emphasized that tension headache results from psychological stress and that stress responses are determined by cognitions about an event or situation. Several concrete examples were provided to illustrate the variety of events that can be perceived as stressful by different individuals and the way in which cognitions can induce psychological stress and headache. In addition, unreasonable expectations (that one should be perfect or liked by everyone) were discussed, and the manner in which these expectations predispose individuals to experience stress was illustrated.

Following presentation of the treatment rationale each group member constructed a list of stressful situations. The therapist, working in turn with each group member, focused on identifying (a) the cues that trigger tension and anxiety, (b) how the client responded when anxious (withdrawal, passivity, etc.), (c) the clients' thoughts prior to becoming aware of tension while tense and subsequently, and (d) the way in which these cognitions appeared to contribute to the clients' tension and headache. Clients were encouraged to learn from group members who proved most adept at this cognitive analysis and to assist other group members in identifying cognitive components of their distress.

As soon as clients became fluent at verbalizing cognitions associated with feelings of distress, they were instructed to deliberately interrupt the sequence of covert events preceding their emotional response at the earliest possible moment. To do this clients were instructed to use signs of impending distress as a signal to engage in cognitive strategies incompatible with the further occurrence of cognitive stress responses. The therapist verbally modeled (Kazdin, 1973) strategies that were designed

¹ Average weekly headache activity scores (HA) were computed by the following formula: $HA = (I \times D)$, where I is intensity of headache and D is the hours of duration of headache. This index of headache is considered to be the most useful measure of headache activity, as it incorporates two separate dimensions of each reported headache.

to enable clients to use each of the three main types of intrapsychic coping responses that have been identified by Lazarus and his co-workers (Lazarus, Averill, & Opton, 1974): cognitive reappraisal, attention deployment, and fantasy. Clients were encouraged to practice these coping skills on a daily basis and to implement these cognitive coping skills at the first sign of headache following the third treatment session.

Combined cognitive and relaxation self-control ($n=10$). This treatment focused on altering both cognitive and muscle contraction responses to stressful situations. The treatment format was similar to that used in the cognitive self-control groups except that muscle relaxation was also taught as a self-control skill (Goldfried & Trier, 1974) for coping with psychological stress and tension headache.

The rationale for this treatment emphasized that tension headache resulted from both psychological stress and sustained muscle contraction responses. Clients were encouraged to monitor both their cognitive responses to stress and their perceived level of muscle tension in stressful situations. In addition to the cognitive coping strategies provided to clients in the cognitive self-control group, clients were taught to use muscle relaxation as a coping skill. Relaxation training followed procedures outlined by Bernstein and Borkovec (1973), with particular attention to facial, neck, and forehead muscles, which are thought to be associated with tension headaches. Approximately equal portions of each session were spent on teaching cognitive and relaxation coping skills. Clients were encouraged to practice muscle relaxation at home on a daily basis and to implement these coping skills at the first sign of headache following the third treatment session.

Headache discussion. This treatment focused on a discussion of the historical roots of symptoms rather than on the development of specific coping skills. Although clients were taught to monitor their cognitive responses to stress in the same manner as they were in the other two treatment groups, no strategies for coping with stress were provided.

The rationale presented for this treatment also described headaches as resulting from psychological stress but emphasized that feelings of distress would improve if clients understood the underlying source of their problems. Group members were encouraged to examine the thoughts and feelings that accompanied their headaches for clues that might be provided to the underlying source of their symptoms. The basic procedures followed by the therapist were designed to increase the clients' self-confidence and self-esteem and to provide a reasonable explanation for the clients' distress in terms of historical events in their lives. The therapist encouraged clients to openly discuss and explore their emotional responses to stressful life events, emphasized similarities among the problems and reactions of group members, and, where possible, offered plausible interpretations designed to link previous life events with current emotional reactions and problems. ("Your anxiety appears to be a natural reaction to the way you

were treated as a child.") The therapists deliberately avoided suggesting specific methods for coping with stress and attempted to prevent other group members from offering this type of advice.

Symptom-monitoring control ($n=10$). Participants assigned to this group recorded their headaches in the same manner as other participants in the study but were informed that due to the large number of applicants, treatment would not be available until a later date. Participants assigned to this group returned their headache data cards by mail on a weekly basis. To maximize compliance with this procedure, they were paid \$10 for accurately recording their headaches.

Six of the initial 39 participants were eliminated prior to data analysis. Two participants withdrew from treatment: 1 participant in the combined group reported that she was uncomfortable participating with another member of her group whose child she taught at school; another participant in the headache discussion group reported that the self-monitoring was exacerbating his headaches. In addition, schedule conflicts prevented 2 participants in the combined group and 1 participant in the headache discussion group from attending a minimum of three sessions. One participant in the symptom-monitoring control group also reported that she lost her headache data cards. Follow-up data were unavailable for 1 participant in the headache discussion group who moved out of state.

Results

Examination of pretreatment scores, presented in Table 1, reveals occasional differences among groups. Although separate analyses of variance revealed that none of these differences were significant, analysis of covariance (with pretreatment scores as the covariate) was used to provide the most accurate assessment of treatment effects.

Headache Recording

Average weekly headache activity scores are presented in Figure 1. Although both self-control groups and the headache discussion group showed substantial reductions in headache activity that were maintained at follow-up, the symptom-monitoring control group showed essentially no change in headache activity.

Analysis of covariance revealed highly significant treatment effects at both posttreatment and follow-up assessments, $F(3, 28) = 7.1$, and $F(3, 27) = 11.8$, respectively, both $ps < .001$. t tests for correlated means re-

Table 1

Means of Pretreatment, Posttreatment, and Follow-Up Evaluations for Major Dependent Variables

Variable	Group			
	Cognitive self-control	Combined self-control	Headache discussion	Headache-monitoring control
Headache activity				
Pre	127.5	183.4	132.4	152.1
Post ^a	35.7	57.2	61.9	147.9
Follow-up ^a	34.1	39.4	47.8	148.1
Headache frequency				
Pre	5.2	5.5	4.6	5.6
Post ^a	2.2	2.8	3.0	4.9
Follow-up ^a	2.0	3.2	2.6	4.8
Headache duration				
Pre	31.2	44.1	30.4	38.1
Post ^a	14.1	20.2	18.8	36.1
Follow-up ^a	11.4	14.9	17.8	34.2
Headache intensity				
Pre	3.7	3.7	4.5	3.7
Post ^a	2.2	2.4	3.2	3.3
Follow-up ^a	2.8	2.2	1.4	3.3
Psychosomatic symptoms ^b				
Pre	80.2	76.4	75.9	85.1
Post ^a	86.5	78.5	90.9	86.1
Follow-up ^a	86.8	77.7	92.2	86.1
Electromyogram (μ V/min)				
Pre	5.3	3.9	3.5	3.5
Post ^a	4.0	3.0	3.2	—
Credibility ^c				
Session 1	9.5	9.9	8.4	—
Post	8.9	9.9	8.3	—
Follow-up	9.1	9.7	9.0	—

^a Adjusted means.^b Larger scores represent a lower incidence of psychosomatic symptoms.^c Sum of two 5-point scales.

vealed that both self-control groups and the headache discussion group showed significant improvements at posttreatment and follow-up assessments (at least $p < .05$), whereas the symptom-monitoring control group showed no change in headache activity. Duncan's new multiple-range test conducted on adjusted posttest and follow-up means further revealed that the two self-control groups and the headache discussion group differed significantly from the symptom-monitoring control group ($p < .05$) but not from one another.

Pretreatment and adjusted posttreatment and follow-up means for additional component measures of headache (frequency, duration, and intensity) are presented in Table 1. Results from separate analyses of covariance conducted on these measures revealed significant

treatment effects on all three of these measures (at least $p < .05$ at follow-up), with the pattern of results very similar to that reported for the composite headache activity score discussed above.

To examine possible therapist differences in outcome, headache activity scores for participants in the three treatment conditions were subjected to Treatment \times Therapist analyses of covariance. Significant therapist effects were obtained at both posttreatment and follow-up assessments, $F(1, 17) = 6.0$, and $F(1, 16) = 6.1$, both $ps < .025$. At both assessments participants treated by the first author obtained somewhat lower adjusted headache activity scores (posttest $M = 26.9$, follow-up $M = 25.9$) than participants treated by the second author (posttest $M =$

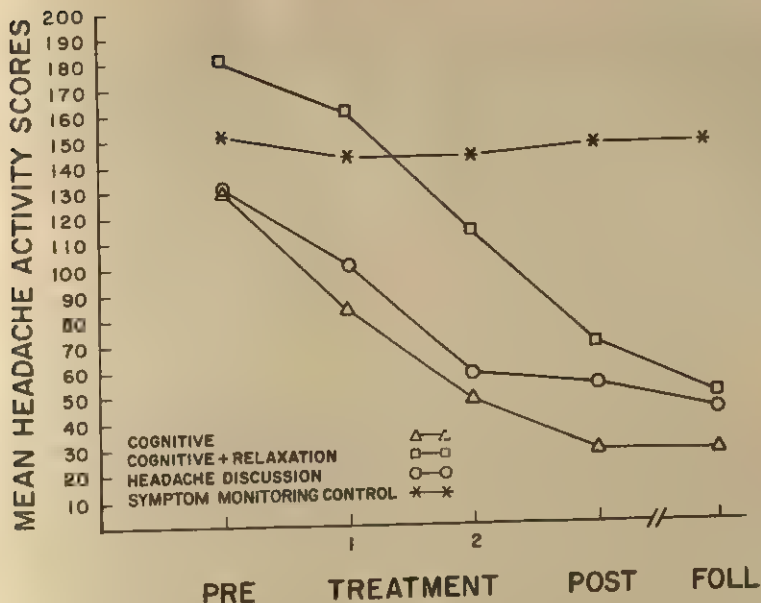


Figure 1. Mean weekly headache activity scores in 2-week blocks. (Foll = follow-up.)

79.9, follow-up $M = 58.9$). However, t tests for correlated means revealed that participants treated by each of the therapists showed significant improvements in headache activity at both assessments (at least $p < .01$). Thus, even though both therapists produced significant reductions in headache activity, reductions by the first author were of a somewhat larger magnitude. Similar analyses conducted on the additional component headache activity scores revealed significant therapist effects on headache duration ($p < .05$) but not on headache frequency or intensity.

Additional Measures

Analysis of covariance conducted on posttest and follow-up psychosomatic checklist frequency scores revealed significant treatment effects at both assessments, $F(3, 28) = 4.2$, $p < .02$, and $F(3, 27) = 4.7$, $ps < .01$, respectively. Although t tests for correlated means revealed that only the cognitive self-control group and the headache discussion group reported fewer psychosomatic symptoms following treatment ($p < .01$), Duncan's test indicated that only the headache discussion and combined cognitive and relaxation

self-control groups differed significantly at posttest ($p < .01$). No therapist differences were observed on this measure. Differences among the four groups in self-esteem approached but did not reach significance, $F(3, 28) = 2.7$, $p < .07$. No differences were observed among the four groups in the number of participants taking medication for their tension headaches, $\chi^2(3) = .28$. However, of the 28 participants taking medication, 17 of 20 participants receiving treatment recorded reductions in the frequency of their medication intake at follow-up, whereas only 1 of 8 participants not receiving treatment recorded reductions (Fisher's exact probability test, $p = .0007$). Correlational analyses revealed no significant relationships between demographic variables, headache history, or self-esteem scores and headache improvement.

Electromyographic Activity

Examination of pretreatment means presented in Table 1 reveals that the severe headache symptoms exhibited by participants in this study were not accompanied by similarly elevated levels of resting frontalis muscle tension. Thus, to the extent that increased mus-

cle tension is associated with headache symptoms exhibited by these individuals, this muscle tension must be elicited by a limited range of, probably, stressful situations. Analyses of covariance revealed that differences in EMG activity among the two self-control groups and the headache discussion group only approached significance at posttest, $F(2, 20) = 2.1, p < .10$. Also, EMG reductions were not significantly correlated with improvements in headache symptoms.

Therapist and Treatment Ratings

It can be seen in Table 1 that both self-control treatments and the headache discussion procedure were rated as highly credible treatments at all three evaluations. However, analyses of variance revealed significant differences in the ratings of these procedures following the first treatment session, $F(2, 29) = 4.9, p < .05$, but not at posttreatment and follow-up assessments. Duncan's test revealed that the headache discussion procedure was rated as slightly less credible than the combined cognitive and relaxation treatment ($p < .05$), whereas the cognitive treatment did not differ from either of the other procedures. The two therapists were rated as equally warm, empathetic, and skillful by participants in their respective groups. However, neither therapist ratings nor treatment ratings were significantly correlated with headache improvement.

Discussion

Results from the present study complement those obtained by Holroyd et al. (1977) in showing that cognitive self-control procedures can provide an effective treatment for chronic tension headache whether they are administered individually or in a group. The group administration used in the present study resulted in reductions in headache activity that were comparable to those obtained when the treatment was individually administered in Holroyd et al. This suggests that therapist time might be effectively conserved by the group administration of these procedures. In addition, these results add to a growing body of evidence (Beck, 1976; Goldfried, 1977;

Mahoney & Arnkoff, in press; Meichenbaum, 1977) supporting the effectiveness of therapeutic procedures designed to alter clients' cognitions in the treatment of anxiety and stress-related disorders.

However, additional findings from the present study raise questions about the effective procedural ingredients of such cognitive therapeutic procedures. Neither the elimination of cognitive coping strategies from the cognitive treatment used in the present study nor the addition of self-control relaxation to this treatment altered its effectiveness. Thus, the positive outcomes that were obtained with this treatment do not appear to have resulted from the specific cognitive coping strategies that were provided. Somewhat similar results have been obtained by Thorpe, Amatu, Blakey, and Burns (1976) who found that not only was the effectiveness of rational emotive therapy not enhanced by the addition of specific self-instructional coping strategies, but the combined procedure was less effective than rational emotive therapy alone on some measures.

Participants in the two self-control treatment groups and in the headache discussion group were interviewed following treatment to obtain additional information concerning the methods that they used to control their headaches. Although all participants in the two self-control groups reported using the self-control procedures that they were taught during treatment, it is of note that all but one of the participants in the headache discussion group also reported devising cognitive self-control procedures for coping with their tension headaches. These strategies appeared to distract the users from worrisome thoughts and/or to enable them to reevaluate the stressor situation. Although the reported methods for controlling headaches were often strikingly similar to those used by participants in the two self-control groups, several individuals developed somewhat unusual methods for managing their headaches. For example, one woman began praying when she noted cognitive symptoms of distress, and she indicated that a brief period of prayer enabled her to approach previously stressful situations with some detachment. A man imaginably engaged in karate exercises. He

reported that these imaginary exercises distracted him from his worries and allowed him to "get a handle" on himself. The only participant who did not report engaging in specific cognitive strategies for coping with headache also showed only minimal improvement in headache activity. These results suggest that the improvements shown by participants in the headache discussion group may have resulted from their use of cognitive coping strategies of their own devising.

Results from other studies that have compared the combination of cognitive therapy and relaxation training procedures with cognitive procedures alone have obtained results consistent with the present findings, indicating that the addition of relaxation training to cognitive interventions does not enhance their effectiveness. Thus, even though combined treatments have been found to be less effective than cognitive interventions (Holroyd, 1976; Meichenbaum et al., 1971; Kanter & Goldfried, Note 1) and equally effective as cognitive interventions (Novaco, 1975; Osarchuk, 1977), they have not been shown to be more effective than cognitive interventions alone. A number of investigators have explained these findings by assuming that clients are unable to master both of these techniques in the brief treatment time typically allowed in these studies (Goldfried, 1977; Meichenbaum et al., 1971). Although frontalis EMG activity was not assessed during relaxation training, the fact that participants in the combined treatment group did not show significant reductions in EMG level following treatment suggests that these individuals may not have adequately mastered the relaxation training procedure. On the other hand, the posttreatment EMG levels of participants in this group were comparable to the post-treatment EMG levels achieved by participants in other studies using relaxation training procedures (Budzynski et al., 1973; Haynes et al., 1975; Hutchings & Reinking, 1976). Thus, the failure to obtain reductions in frontalis muscle activity may have resulted from the initially low resting EMG levels exhibited by participants in this study.

Frontalis muscle activity was not significantly associated with headache symptoms prior to treatment, and reductions in EMG

activity were not correlated with headache improvement following treatment. This weak association between resting levels of frontalis EMG activity and tension headache has been found in a number of recent studies (Cox et al., 1975; Epstein & Abel, 1977; Haynes et al., 1975; Holroyd et al., 1977). It seems likely that muscle contraction responses to specific stressful situations might contribute to tension headaches even though these responses might not be elicited in the relaxed laboratory assessment situation. Therefore, research on tension headache would probably benefit from the development of methods for assessing such responses to laboratory or real-life stressors.

The fact that both the self-control and headache discussion treatments yielded similar outcomes suggests that elements common to these interventions may have accounted for their results. For example, the provision of a causal explanation for distressing symptoms may have served to increase clients' belief in their ability to cope with their symptoms (Frank, 1974; Murray & Jacobson, *in press*). Although somewhat different explanations of the therapeutic process were provided in each of the groups, they all emphasized that clients could master their symptoms. Thus, specific cognitive distortions or muscle contraction responses eliciting tension headache could be combated by engaging in specific coping responses or by understanding their historical antecedents, which were no longer present and therefore need not influence current responses. Such explanations, which increase an individual's belief in their ability to cope with previously debilitating symptoms, might be expected to lead to a greater initiation and persistence of coping behavior (Bandura, 1977). Similarly, training in identifying a cognitive component of distress may have facilitated control of tension headaches by sensitizing clients to early signs of psychological stress, thus providing cues for appropriately engaging in coping responses (Meichenbaum, 1975). It may be less crucial to provide clients with specific coping responses than to insure that they monitor the insidious onset of symptoms and are capable of engaging in some sort of cognitive or behavioral response incompatible

with the further exacerbation of symptoms. Clearly, further research will be required to determine the way that these factors contribute to therapeutic change.

Although clients seen by each of the therapists showed substantial reductions in headache activity, somewhat larger reductions in headache activity were obtained by one of the therapists. Since no differences in outcome were observed when similar treatments were ministered individually in a previous study (Holroyd et al., 1977), the differences obtained in the present study may have resulted from therapist differences in experience with group therapy procedures.

Reference Notes

1. Kanter, N. J., & Goldfried, M. R. *Relative effectiveness of rational restructuring and self-control desensitization in the reduction of interpersonal anxiety*. Unpublished manuscript, State University of New York at Stony Brook, 1976.
2. Linehan, M. M., & Goldfried, M. R. *Assertion training for women: A comparison of behavioral rehearsal and cognitive restructuring*. Paper presented at the ninth annual convention of the Association for Advancement of Behavior Therapy, San Francisco, December 1975.

References

- Ad Hoc Committee on Classification of Headache. Classification of headache. *Journal of the American Medical Association*, 1962, 179, 717-718.
- Bakal, D. Headache: A biopsychological perspective. *Psychological Bulletin*, 1975, 82, 369-382.
- Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 1977, 84, 191-215.
- Beck, A. T. *Cognitive therapy and the emotional disorders*. New York: International Universities Press, 1976.
- Bernstein, D. A., & Borkovec, T. D. *Progressive relaxation training: A manual for helping professionals*. Chicago: Research Press, 1973.
- Budzynski, T. H., Stoyva, J. M., Adler, C. S., & Mullaney, D. J. EMG biofeedback and tension headache: A controlled outcome study. *Psychosomatic Medicine*, 1973, 35, 484-496.
- Cox, D. J., Freundlich, A., & Meyer, R. G. Differential effectiveness of electromyograph feedback, verbal relaxation instructions, and medication placebo with tension headaches. *Journal of Consulting and Clinical Psychology*, 1975, 43, 892-898.
- Di Loreto, A. O. *Comparative psychotherapy: An experimental analysis*. Chicago: Aldine-Atherton, 1971.
- Epstein, L. H., & Abel, G. G. An analysis of biofeedback training effects for tension headache patients. *Behavior Therapy*, 1977, 8, 37-47.
- Frank, J. D. *Persuasion and healing: A comparative study of psychotherapy*. New York: Schocken Books, 1974.
- Goldfried, M. R. The use of relaxation and cognitive relabeling as coping skills. In K. B. Stuart (Ed.), *Behavioral self-management: Strategies and outcomes*. New York: Brunner/Mazel, 1977.
- Goldfried, M. R., Decentecio, E. T., & Weinberg, L. Systematic rational restructuring as a self-control technique. *Behavior Therapy*, 1974, 5, 247-254.
- Goldstein, A. P., Heller, K., & Sechrest, L. B. *Psychotherapy and the psychology of behavior change*. New York: Wiley, 1966.
- Goldfried, M. R., Linehan, M. M., & Smith, J. L. Reduction of test anxiety through cognitive restructuring. *Journal of Consulting and Clinical Psychology*, 1978, 46, 32-39.
- Goldfried, M. R., & Trier, C. S. Effectiveness of relaxation as an active coping skill. *Journal of Abnormal Psychology*, 1974, 83, 348-355.
- Haynes, S., Griffin, P., Mooney, D., & Parise, M. Electromyographic biofeedback and relaxation in the treatment of muscle contraction headache. *Behavior therapy*, 1975, 6, 672-678.
- Holroyd, K. Cognition and desensitization in the group treatment of test anxiety. *Journal of Consulting and Clinical Psychology*, 1976, 44, 991-1001.
- Holroyd, K., Andrasik, F., & Westbrook, T. Cognitive control of tension headache. *Cognitive Therapy and Research*, 1977, 1, 121-133.
- Hutchings, D. F., & Reinking, R. H. Tension headaches, what form of therapy is most effective? *Biofeedback and Self-Regulation*, 1976, 1, 183-190.
- Kashiwagi, T., McClure, J. N., & Wetzel, R. D. Headache and psychiatric disorders. *Diseases of the Nervous System*, 1972, 33, 659-663.
- Kazdin, A. E. Covert modeling and the reduction of avoidance behavior. *Journal of Abnormal Psychology*, 1973, 81, 87-95.
- Langer, E. J., Janis, I. L., & Wolfer, J. A. Reduction of psychological stress in surgical patients. *Journal of Experimental Social Psychology*, 1975, 11, 155-165.
- Lazarus, R., Averill, J., & Opton, E. The psychology of coping: Issues of research and assessment. In G. Coelho, D. Hamburg, & J. Adams (Eds.), *Coping and adaption*. New York: Basic Books, 1974.
- Mahoney, M. J., & Arnkoff, D. B. Cognitive and self-control therapies. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (2nd Ed.). New York: Wiley, in press.
- Martin, M. J. Muscle-contraction headache. *Psychosomatics*, 1972, 13, 16-19.
- Meichenbaum, D. Cognitive modification of test anxious college students. *Journal of Consulting and Clinical Psychology*, 1972, 39, 370-380.
- Meichenbaum, D. Toward a cognitive theory of self-control. In G. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research* (Vol. 1). New York: Plenum Press, 1975.

- Meichenbaum, D. *Cognitive-behavior modification*. New York: Plenum Press, 1977.
- Meichenbaum, D. H., Gilmore, J. B., & Fedoravichus, A. Group insight versus group desensitization in treating speech anxiety. *Journal of Consulting and Clinical Psychology*, 1971, 36, 410-421.
- Meichenbaum, D. H., Turk, D., & Burnstein, S. The nature of coping with stress. In I. Sarason & C. Spielberger (Eds.), *Stress and anxiety* (Vol. 2). Washington, D.C.: Hemisphere, 1975.
- Moleski, R., & Tosi, D. J. Comparative psychotherapy: Rational-emotive therapy versus systematic desensitization in the treatment of stuttering. *Journal of Consulting and Clinical Psychology*, 1976, 44, 309-311.
- Murray, E. J., & Jacobson, L. I. Cognition and learning in traditional and behavioral therapy. In S. L. Garfield & A. E. Bergen (Eds.), *Handbook of psychotherapy and behavior change* (2nd ed.). New York: Wiley, in press.
- Novaco, R. *Anger control: The development and evaluation of an experimental treatment*. Lexington, Mass.: Heath, 1975.
- Osarchuk, M. A comparison of a cognitive, a behavior therapy and a cognitive plus behavior therapy treatment of depressed outpatients. *Cognitive Therapy and Research*, 1977, 1, 17-37.
- Rush, A. J., Beck, A. T., Kovacs, M., & Hollon, S. Comparative efficacy of cognitive therapy and pharmacotherapy in the treatment of depressed outpatients. *Cognitive Therapy and Research*, 1977, 1, 17-37.
- Ryan, V. L., Krall, C. A., & Hodges, W. F. Self-concept change in behavior modification. *Journal of Consulting and Clinical Psychology*, 1976, 44, 638-645.
- Steinmark, S., & Borkovec, T. Active and placebo treatment effects on moderate insomnia under counterdemand and positive demand instructions. *Journal of Abnormal Psychology*, 1974, 83, 157-163.
- Thorpe, G. L. Desensitization, behavior rehearsal, self-instructional training and placebo effects on assertive-refusal behavior. *European Journal of Behavioral Analysis and Modification*, 1975, 1, 30-44.
- Thorpe, G. L., Amatu, H. I., Blakey, R. S., & Burns, J. E. Contributions of overt instructional rehearsal and "specific insight" to the effectiveness of self-instructional training: A preliminary study. *Behavior Therapy*, 1976, 7, 504-518.
- Wolff, H. G. *Headache and other head pain*. New York: Oxford University Press, 1963.
- Wolfe, J. L., & Fodor, I. G. Modifying assertive behavior in women: A comparison of three approaches. *Behavior Therapy*, 1977, 8, 567-574.

Received August 8, 1977 ■

Comparison of Linear and Configural MMPI Diagnostic Methods With an Uncontaminated Criterion

Ronald A. Giannetti, James H. Johnson, Daniel E. Klingler,
and Thomas A. Williams

Department of Psychiatry, University of Utah

Previous studies of linear and configural Minnesota Multiphasic Personality Inventory (MMPI) diagnostic predictors have suffered from varying degrees of criterion contamination. We replicated and extended previous findings with 572 subjects who had been diagnostically classified without MMPI contamination. The Goldberg linear equations derived from MMPI group profiles achieved 84% accuracy in classifying group profiles and a 14% increment over base-rate accuracy in classifying individual profiles. The Goldberg linear equation and several configural methods for discriminating psychotics from neurotics were compared. The linear equation was found to be most accurate. Conflicting results in previous articles suggest that criterion contamination must be avoided in prediction studies. A possible use for the group profile classification equations in evaluating experimental studies is suggested.

Several configural rules and linear indices have been developed to diagnostically classify Minnesota Multiphasic Personality Inventory (MMPI) profiles. Peterson (1954) specified a number of signs that discriminated patients who later become schizophrenic. These signs are (a) *T* scores on four or more clinical scales greater than 70; (b) *F* greater than 65; (c) *Sc* greater than *Pt*; (d) *Pa* or *Ma* greater than 70; (e) *Pa*, *Sc*, or *Ma* greater than *Hs*, *D*, and *Hy*; and (f) *D* greater than both *Hs* and *Hy*. Taulbee and Sisson (1957) developed rules to discriminate the profiles of schizophrenics from those of neurotics. Their method compares 16 scale pairs for elevation and tallies the number of pairs that differ in the scored direction. Experimental results have shown this method to be more accurate than the judgment of clinicians. Meehl and Dahlstrom (1960) constructed and cross-validated a highly complex, multistep, configural algorithm for discriminating the profiles of psychotics from neurotics. They elucidated the central assumption of the configural approach, namely, that the infor-

mation required to make discriminations resides primarily in nonlinear relationships among MMPI scales.

Goldberg (1965, 1969) challenged this assumption in two comprehensive studies that compared numerous techniques for discriminating psychotic from neurotic MMPI profiles. His results indicated that a five-variable linear composite of MMPI scales ($L + Pa + Sc - Hy - Pt$) was superior to previously developed configural rules, profile typologies, actuarial tables, several nonlinear (e.g., Bayesian) actuarial techniques, and clinicians' judgments in making the psychotic/neurotic discrimination.

Goldberg (1972) also applied the linear model to group profiles. He hypothesized that his procedure results in indicators of underlying processes that are more useful than results derived from individual profiles. His rationale was that the process of averaging individual profiles reduces the error, or "noise," inherent in individual profiles and results in a clearer "signal" of underlying processes. Goldberg classified Lanyon's (1968) compilation of MMPI group profiles into the molar diagnostic categories of psychotic, neurotic, sociopathic, and normal. He then

Requests for reprints should be sent to Ronald A. Giannetti, Department of Psychiatry and Behavioral Science, Box 1980, Norfolk, Virginia 23501.

developed a set of three linear equations, which are applied sequentially to group profiles for diagnostic sorting. The first equation separates normals from psychiatric groups. The second separates sociopathic from psychiatric groups. The final equation separates the remaining groups into psychotic and neurotic. This procedure classified profiles with remarkable accuracy (93%-99%). In addition, the linear equation, earlier derived from individual profiles to discriminate psychotics from neurotics, was replicated for group profiles. These results provide support for the use of linear rather than configural systems for MMPI profile classification.

However, the issue of configural rules versus linear combinations has not been conclusively resolved. In each of the studies cited above, there was some criterion contamination; the MMPI results were used in varying degrees to determine the criterion diagnoses that were being predicted. Goldberg (1965) was aware of this difficulty and reanalyzed his individual profile data after dividing his sample into groups considered to be "least contaminated" and "most contaminated." The results of these analyses suggested that criterion contamination may have had significant effects on the predictive validities. In the least contaminated sample, there was no difference between the predictive validities of the Goldberg equation and the Meehl-Dahlstrom rules. He also found that if one assigns a rank order to scores by elevation to the eight psychopathology scales, the sum of the ranks of *D*, *Hy*, and *Pt* (the most elevated scale receives a rank score of 1; low scores are considered neurotic) has the highest validity coefficient in the least contaminated sample. Thus, although the preponderance of his results favored the linear equation, one configural method equaled and another outperformed it.

In a replication study, Goodson and King (1976) compared the predictive accuracy of the Peterson signs and the Goldberg equation. In contradiction to Goldberg's results, they found the Peterson signs to be superior in two samples. However, the criterion diagnoses in one sample were based entirely on the MMPI and, hence, were contaminated. Their findings were unclear as to whether the MMPI contributed to the diagnoses in the second sample.

The present study was undertaken to replicate and extend certain of the results of the above studies using an uncontaminated criterion. The purposes of this study were to (a) attempt to replicate Goldberg's (1972) results with group profiles; (b) estimate the diagnostic classificatory accuracy of the equations derived from group profiles when applied to individual profiles; and (c) compare the relative classificatory accuracy of the various "cookbook" methods for discriminating psychotic from neurotic profiles (specifically, the Goldberg equation, the Meehl-Dahlstrom rules, the sum-of-ranks rule, the Taulbee-Sission signs, and the Peterson signs).

Method

Subjects

The sample was composed of male veterans who were evaluated at the time of application for treatment at the Psychiatric Assessment Unit of the Salt Lake City Veterans Administration Hospital. Subject selection was limited to those veterans who completed the experimental evaluation and testing within 48 hours of referral. In addition, all subjects whose MMPI had more than 30 missing items or had an *F-K* dissimulation index (Gough, 1950) greater than 14 were eliminated. This resulted in a sample of 572 subjects classified into seven psychotic groups ($n = 218$), three sociopathic groups ($n = 208$), three neurotic groups ($n = 88$), and one "no mental illness" group ($n = 58$). The mean age for the sample was 37.7 years ($SD = 12.4$). The mean IQ estimated from the Shipley-Hartford (Paulson & Lin, 1970) was 103.4 ($SD = 12.7$).

Instrumentation

The Current and Past Psychopathology Scales (CAPPS) structured recording form for evaluating current and past psychopathology and social functioning (Endicott & Spitzer, 1972) was used as the diagnostic data base form. The CAPPS consists of 41 items about current psychopathology and 130 items about past psychopathology, personality characteristics, and academic, occupational, and interpersonal adjustment. It is a rationally constructed instrument that was developed to provide coverage of symptoms generally considered to be important in the evaluation of diagnosis, severity of illness, and prognosis. Endicott and Spitzer have shown that interjudge reliabilities for individual items are very high (range = .68-1.00). For purposes of this experiment, the CAPPS was used as data input to the DIAGNO II diagnostic program (Spitzer & Endicott, 1969). The DIAGNO II program was rationally constructed to mimic the clinician's diagnostic reasoning, given ratings on the CAPPS items as the source of input information. This program has

been shown to produce diagnoses that agree with standard clinical diagnoses as well as experienced clinicians agree with each other and at the same time to reduce situational diagnostic bias (Spitzer & Endicott, 1969).

Subjects were also administered the booklet form of the MMPI and the Shipley-Hartford.

Procedure

CAPPS interviews were administered by three staff members (a clinical psychologist, a social worker, and a nurse) who had been trained using materials provided by Jean Endicott. The order of administration of the CAPPS and MMPI was dependent on the availability of an interviewer. In those instances in which the MMPI was administered prior to the CAPPS, the interviewer had no knowledge of the MMPI results.

Although previous research with the CAPPS has shown that well-trained interviewers produce highly reliable item ratings, the interjudge diagnostic reliability of the staff members was studied. An additional sample of 30 male applicants for psychiatric care served as subjects for this study. Three interviewers recorded CAPPS ratings for each of the subjects. These ratings were input to DIAGNO II. Output consisted of 30 sets of three diagnoses made using the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1968). The kappa technique, developed by Fleiss (1971) for measuring nominal agreement among raters, was used to compute interjudge diagnostic reliability. The value of kappa was .45 for all diagnoses for the 30 subjects. This value was significantly greater than zero ($p < .001$) and somewhat greater than the mean of the mean of kappas ($\kappa = .40$) reported by Spitzer and Fleiss (1974) for 12 specific diagnoses in their review of seven diagnostic reliability studies.

A previous study on our patient population showed that experienced senior clinicians agree with DIAGNO-II diagnoses no better than with the treatment clinicians' diagnoses (Klingler, Miller, Johnson, & Williams, 1977).

To cross-validate the predictors of group profiles, a mean profile was calculated for each of the diagnostic groups. Group profiles were then scored on the relevant predictors derived in Goldberg's (1972) study (normal vs. psychiatric, $Hs + 2Pd - Ma$; psychiatric vs. sociopathic, $2Pd - Hy - Sc$; and psychotic vs. neurotic, $L + Pa + Sc - Hy - Pt$). The means and standard deviations of these scores were calculated over the number of mean profiles in the psychotic, sociopathic, and neurotic categories. Only the mean could be calculated for the no mental illness group.

The diagnostic accuracy of the equations derived for the group profiles was then tested on individual profiles. Since these equations are applied sequentially, and since we lacked information about the expected distribution of group differences to set cutting scores, the following steps were taken: First, the mean and standard deviation for each equation were calculated for the portion of the sample to which they would be

applied. Thus, for the "normal versus psychiatric" equation, these statistics were calculated over the entire sample. For the "psychiatric versus sociopathic" equation, these calculations were made with the no mental illness group removed, and so forth. Second, assuming a normal distribution of scores on the predictors, the cutting scores for classification were set such that the selection ratio would equal the diagnostic base rates when each equation was applied.

The psychotic versus neurotic predictors were subjected to two analyses for comparability with Goldberg's (1965) results. Because the Meehl-Dahlstrom rules and Taulbee-Sisson signs permit an indeterminate classification, the Goldberg equation and the sum-of-ranks rule were also permitted an indeterminate classification. Using Goldberg's cutting scores, the indeterminate ranges were designated as scores between 40 and 49 on the Goldberg equation and scores of 11 or 12 on the sum-of-ranks rule.

A second analysis was performed to force a prediction for each case, because an indeterminate range has not been designated for the Peterson signs. In this analysis only the Goldberg equation, the sum-of-ranks rule, and the Peterson signs could be compared. The cutting points for predicting psychosis were scores greater than 45 for the Goldberg equation, greater than 11 for the sum-of-ranks rule, and more than 2 signs present for the Peterson signs.

One further procedure was necessary to insure comparability to Goldberg's results. His analyses were based on a sample that approached a 50% split between the criterion diagnoses. The present sample contained far more psychotics than neurotics. Thus, all 88 individuals with neurotic diagnoses were included in the analysis. An equal number of psychotics was selected by randomly sampling from each of the seven psychotic diagnoses in proportion to the number of individuals receiving each diagnosis.

Results

Analysis of the group profiles confirms Goldberg's results. Table 1 displays the means and standard deviations obtained on the group profile predictors and those obtained in Goldberg's (1972) derivation sample. Despite the small sample of groups in this study, the separations among the groups were very large and in the expected directions. The poorest separation was between psychotics and neurotics; yet, the means were more than two standard deviations apart, one-tailed $t(8) = 2.95$, $p < .01$. Since a test of significance does not necessarily imply a strong degree of statistical association (Hays, 1963), it was decided to calculate the omega-square for this contrast. The value of ω^2 was .44, suggesting a strong degree of discrimination between groups. Further evidence for the strength of

Table 1
Comparison of the Results of the Goldberg Group-Profile Classification Formulae with a Replication Sample

Predictor	Present study						Goldberg (1972)					
	Psychotic		Neurotic		Sociopathic		Psychotic		Neurotic		Sociopathic	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Psychiatrists vs. normal ($H_s + 2Pd - Ma$)	149	8	152	9	149	5	136	14	141	7	140	9
Sociopathic vs. psychiatric ($2Pd - Hy - Sc$)	-2	5	-3	10	18	3	-2	8	-10	9	24	8
Psychotic vs. neurotic ($L + Pa + Sc - Hy - Pt$)	54	6	41	5	3		67	12	37	7	41	
No. group profiles	7		3		3		22		19		78	

Table 2

Valid Positive Predictions for Individual Profiles Using Group-Profile-Derived Predictors

Category	Valid positives (%)	n
Schizophrenia, paranoid type	59	22
Other major affective disorder	58	19
Psychotic depressive reaction	50	36
Schizophrenia, catatonic type	50	16
Schizophrenia, chronic undifferentiated type	44	84
Acute schizophrenic episode	39	23
Schizophrenia, schizoaffective type	28	18
Total psychotic	46	218
Drug dependence	66	44
Unspecified alcoholism	52	130
Antisocial personality	50	34
Total sociopathic	54	208
Hysterical neurosis	50	22
Depressive neurosis	25	32
Anxiety neurosis	19	34
Total neurotic	36	88
No mental illness	19	58
Total sample	45	572

these results was obtained by applying the cutting scores for the classification equations suggested in the derivation study to this sample. Only 2 of the 14 group profiles were misclassified: "depressive neurosis" and "no mental disorder." This represents 86% accuracy of classification in a small replication sample.

These equations were then used to sequentially classify the 572 individual profiles. After each equation was applied, the fourfold hit-miss table was analyzed using McNemar's test for correlated proportions (Siegel, 1956). This was done to test the hypothesis that the distribution of predictions was not significantly different from the distribution of diagnoses. Since a hypothesis of no difference was being tested, the significance level was set at .10 to avoid a Type I error. The one-tailed chi-square values ($df = 1$) for the three prediction stages were .01, 1.6, and 2.2, respectively ($p > .10$).

Table 3

Comparison of Four Methods for Discriminating Psychotics From Neurotics Permitting an Indeterminate Classification

Method	Hits		Misses		Indeterminate	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Goldberg equation	95	65.5	50	34.5	31	17.6
Sum of ranks (<i>Hy</i> , <i>D</i> , & <i>Pt</i>)	77	55.8	61	44.2	38	21.6
Meehl-Dahlstrom rules	63	46.7	72	53.3	41	23.3
Taulbee-Sisson signs	45	45.9	53	54.1	78	44.3

Note. Hit-miss percentages are based on the number of profiles classified.

The distributions of predicted and criterion diagnoses were not significantly different.

Table 2 displays the hit rates for each specific diagnosis, each molar diagnostic category, and the total sample. The sequential approach was most accurate in identifying characterological and psychotic syndromes (with the exception of schizoaffective schizophrenia) and least accurate in identifying the no mental illness profiles. Random assignment of profiles to the molar diagnostic categories, given base rates, should produce an expected hit rate of 31%. For all subjects the hit rate was 45%, a 14% improvement over chance accuracy.

Results of the comparison of the various methods for discriminating psychotics from neurotics, with an indeterminate category permitted, are shown in Table 3. These results confirm Goldberg's findings. The Goldberg equation resulted in the most hits, the fewest misses, and left the fewest individuals unclassified. Calculating a phi coefficient on the fourfold hit-miss contingency table on individuals for whom a prediction was made yielded a $\phi = .320$, $\chi^2(1) = 14.85$, $p < .0005$. The sum-of-ranks rule yielded a $\phi = .121$, which did not achieve significance, $\chi^2(1) = 2.01$, $p > .10$. Surprisingly, the Meehl-Dahlstrom rules and the Taulbee-Sisson signs achieved less than chance accuracy, and the latter left nearly 45% of the sample unclassified.

Results with the indeterminate category excluded are shown in Table 4. In this analysis the Goldberg equation again was the most accurate, followed closely by the Peterson signs. The sum-of-ranks rule did not produce a significant result.

Discussion

The results obtained in discriminating psychotics from neurotics support Goldberg's previous results, which suggest the superiority of the linear model over configural methods. However, it is noted that none of the presently available linear, sequential-linear, or configural methods works very well in predicting diagnostic categories from individual profiles. In each case the incremental accuracies obtained above the base rates were quite modest. Even though actuarial methods are accurate in classifying groups, they appear to have little validity for the practical clinical problem of classifying individuals.

The present findings also serve to highlight the potent effects of criterion contamination. In Goldberg's (1965) most contaminated sample, the Meehl-Dahlstrom rules and the Taulbee-Sisson signs achieved the creditable validity coefficients of .42 and .40, respectively. These coefficients fell to .29 and .27 in the least contaminated sample. In the present study both methods resulted in accuracy below chance. In addition, a number of other pre-

Table 4
Comparison of Three Methods for Distinguishing Psychotics From Neurotics, Excluding an Indeterminate Classification

Method	% hits	ϕ	χ^2
Goldberg equation	61.9	.239	10.09*
Peterson signs	60.8	.216	8.23*
Sum of ranks	55.7	.114	2.28

Note. $df = 1$ for chi-squares.

* $p < .005$ ($n = 176$).

dictors in the 1965 study showed the opposite effect on their coefficients; they had lower validities in the more contaminated sample. This suggests that criterion contamination must be scrupulously avoided. If it occurs, even to a slight degree, results obtained from such data can be misleading.

Despite possible difficulties with Goldberg's (1972) group profile derivation sample, this technique appears quite robust on cross-validation. This replication lends greater credence to the argument that group averaging serves to filter out much of the measurement error inherent in individual profiles, revealing a composite variable closely related to the criterion. Unfortunately, the high degree of accuracy in classifying groups does not generalize to classifying individuals.

The accuracy of group classification procedures may be used to advantage in evaluating research about psychopathology. Many studies of pathological groups report MMPI mean scores. Two procedures are commonly used to relate the findings of these studies to the MMPI. The mean profile of a group is interpreted as if it were an individual profile, and the percentage of the most frequently appearing two-point codes in the group are reported. Although these procedures are useful for descriptive purposes, application of the sequential formulae to the group mean profile would provide strong evidence as to which population the group under study was drawn from: normal, sociopathic, neurotic, or psychotic.

This may be illustrated by two examples from the recent literature. Rader (1977) compared the MMPIs of three groups of men arrested for either indecent exposure, assault, or rape. The mean profile of the exposure group was within normal limits, with *Pd* being the most elevated. Rader concluded that the exposure group was comprised primarily of mild nonconformists who occasionally test societal limits. The sequential formulae result in a sociopathic classification for this mean profile. This indicates that the exposure group is more deviant than visual inspection of the mean profile would suggest. In another study, Widom (1977) tested the efficacy of an advertisement for recruiting noninstitutionalized psychopaths for experimental study.

Applying the Goldberg rules to the mean profile of 23 men recruited by this procedure results in a sociopathic classification.¹ This corroborates Widom's findings that a sample of psychopaths had been recruited by this method. Thus, Goldberg's linear combinations are a valuable adjunct for interpreting group results.

¹ MMPI scale scores were estimated visually from a figure in the study.

References

- American Psychiatric Association. *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, D.C.: Author, 1968.
- Endicott, J., & Spitzer, R. L. Current and Past Psychopathology Scales (CAPPS): Rationale, reliability, and validity. *Archives of General Psychiatry*, 1972, 27, 678-687.
- Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 1971, 76, 378-382.
- Goldberg, L. R. Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs*, 1965, 79(9, Whole No. 602).
- Goldberg, L. R. The search for configural relationships in personality assessment: The diagnosis of psychosis vs. neurosis from the MMPI. *Multivariate Behavioral Research*, 1969, 4, 523-536.
- Goldberg, L. R. Man versus mean: The exploitation of group profiles for the construction of diagnostic classification systems. *Journal of Abnormal Psychology*, 1972, 79, 121-131.
- Goodson, J. H., & King, G. D. A clinical and actuarial study on the validity of the Goldberg index for the MMPI. *Journal of Clinical Psychology*, 1976, 32, 328-335.
- Gough, H. G. The *F* minus *K* dissimulation index for the MMPI. *Journal of Consulting Psychology*, 1950, 14, 408-413.
- Hays, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.
- Klingler, D. E., Miller, D. A., Johnson, J. H., & Williams, T. A. Process evaluation of an on-line computer-assisted unit for intake assessment of mental health patients. *Behavior Research Methods and Instrumentation*, 1977, 9, 110-116.
- Lanyon, R. I. *A handbook of MMPI group profiles*. Minneapolis: University of Minnesota Press, 1968.
- Meehl, P. E., & Dahlstrom, W. G. Objective configural rules for discriminating psychotic from neurotic MMPI profiles. *Journal of Consulting Psychology*, 1960, 24, 375-387.
- Paulson, M. J., & Lin T. T. Predicting WAIS IQ from Shipley-Hartford scores. *Journal of Clinical Psychology*, 1970, 26, 454-461.
- Peterson, D. R. The diagnosis of subclinical schizo-

- phrenia. *Journal of Consulting Psychology*, 1954, 18, 198-200.
- Rader, C. M. MMPI profile types of expositors, rapists, and assaulters in a court services population. *Journal of Consulting and Clinical Psychology*, 1977, 45, 61-69.
- Siegel, S. *Non-parametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- Spitzer, R. L., & Endicott, J. DIAGNO II: Further developments in a computer program for psychiatric diagnosis. *American Journal of Psychiatry*, 1969, 125, 12-21.
- Spitzer, R. L., & Fleiss, J. L. A re-analysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry*, 1974, 125, 341-347.
- Taulbee, E. S., & Sisson, B. D. Configurational analysis of MMPI profiles of psychiatric groups. *Journal of Consulting Psychology*, 1957, 21, 413-417.
- Widom, C. S. A methodology for studying noninstitutionalized psychopaths. *Journal of Consulting and Clinical Psychology*, 1977, 45, 674-683.

Received August 9, 1977 ■

Longitudinal Study of Marital Success and Failure

P. M. Bentler and Michael D. Newcomb
University of California, Los Angeles

Personality and background questionnaires were administered to newly married couples. Four years later these couples were followed up to determine their marital status and satisfaction. Our findings indicate that (a) correlational similarity as well as mean differentiation between partners was higher in the still-married group than the divorced group; (b) accuracy of self-perception was marginally reflective of marital success; (c) living together before marriage had no apparent effect on the outcome of marriage; (d) divorced couples appeared to face qualitatively different problems than married couples; and (e) longitudinal prediction of marital adjustment was possible, with prediction based on signed, equal weights yielding $R \cong .70$. It appears that variation in marital outcome is most accurately predicted from personality and not demographic variables, based largely on data from women.

This study attempts to clarify the role of marital partners' personality traits on the success or failure of their marriage. We gathered information on background, personality, and peer assessment of personality variables at the beginning of the marriages in our sample. Four years later we determined the outcome of these marriages in terms of staying together or divorcing and the quality of these marriages using a marital adjustment scale. We then compared the separated or divorced with the still-married couples in terms of the variables that we had assessed at the earlier period. Since our independent variables—personality and background data—were assessed at the beginning of the marriages, we were able to construct several useful regression equations to predict marital adjustment.

There have been only a few longitudinal

studies addressed to the question of what draws people together as couples and what keeps them there once joined. Several of these studies focus on factors that contribute most toward a couple deciding to marry or to separate before marriage (Burgess & Wallin, 1953; Hill, Rubin, & Peplau, 1976; Udry, 1967). This study addresses a different segment in the marriage process, beginning with the marriage itself. We wanted to know how and which personality traits, brought to the marriage by each partner, affected the subsequent outcome and quality of that marriage 4 years later. This is an old problem; Kelly (1939) and Terman and Oden (1947), both using longitudinal designs, addressed a similar question. Their assessment of personality was in the form of a general personality factor that could not be broken down into identifiable or logical components. They were looking for the overall contribution of "personality," as a global concept, to the outcome of marriage. The present study has assessed a wide variety of heterogeneous personality traits in order to specify personality to a much greater extent. Burgess and Wallin (1953) have also studied engagement and marital adjustment longitudinally. They eliminated divorced or separated couples from their sample, looked only at intact marriages, and used the Thurstone Personality Inventory, designed to as-

This report was prepared for the International Conference on Love and Attraction, University College of Swansea, Wales, September 1977. The investigation was supported in part by a Research Scientist Development Award (K02-DA00017) and Research Grant DA01070 from the U.S. Public Health Service. The assistance of Bonnie Barron, Leslie Hutcherson, Marion Hee, and George Speckart is gratefully acknowledged.

Requests for reprints should be sent to P. M. Bentler, Department of Psychology, University of California, Los Angeles, California 90024.

sess primarily neurotic problems, which, as Tharp (1963) pointed out, makes the results less applicable to a nonpathological population.

Trait research has also generated many cross-sectional studies of married and divorced groups. Even though personality traits included in recent research have focused more on those that are specifiable and nonneurotic, the question of causation or prediction has remained clouded and unanswered in cross-sectional designs (DeYoung & Fleischer, 1976; Locke, 1951; Murstein & Glaudin, 1966; Pickford, Signori, & Rempel, 1966a, 1966b; Singh, Nigam, & Saxena, 1976). Whereas the causative effects of background characteristics can reasonably be studied cross-sectionally in that they are not subject to change by the passing of time or the pressures of marriage, this cannot necessarily be said for personality traits. If one studies a marriage failure cross-sectionally, one does not know whether the personality pattern was the cause or consequence of marital breakup.

The current research has attempted to avoid these pitfalls by longitudinally studying a wide variety of nonpathological personality traits and the effects that they have on marital outcome. Since the nature of our study does not permit the use of initial random assignment nor the manipulation of personality variables, it represents something less than the ideal longitudinal experiment. Yet, the design would seem to be quite functional within these constraints. Our general hypotheses were that (a) those marriages that are still intact will have shown, at the beginning of their marriage, more similarity (homogamy) of partners on personality traits and background items than the separated or divorced couples (Barton & Cattell, 1972; Cattell & Nesselroade, 1967; DeYoung & Fleischer, 1976; Kernodle, 1969; Murstein, 1961, 1967; Singh et al., 1976; Holz, Note 1); (b) there will be a greater peer-self agreement on personality for the still-married couples than for the split marriages (Murstein & Beck, 1972; Weigel, Weigel, & Richardson, 1973); (c) living together before marriage will increase the probability of marital success; (d) problems of a married

couple will be qualitatively different than problems of a divorced couple; and (e) significant longitudinal prediction of marital success based on traits and demographics is possible.

Method

Sample

Our original sample consisted of 162 newly married couples solicited from the rolls of the local marriage license bureau. After follow-up 4 years later, our working sample consisted of 77 couples: 53 were still married, and 24 had separated or divorced. The attrition in sample size was primarily due to the inability to contact many couples, as well as death in two instances. At the time of follow-up, we received responses from 89% of the deliverable letters, which seems reasonably high. Los Angeles is a highly mobile community, which made recontacting the original sample extremely difficult. Only a very small percentage of our original couples lived at the same address 4 years after initial contact with them. Nonetheless, we do not feel that our sample reduction has been caused by any systematic bias; it represents a fair sample of couples under realistic field conditions. All studies of this type have problems with sampling bias (e.g., volunteer bias, motivation) that we feel are unavoidable. This bias should not invalidate the results, but it stresses the importance of replication.

All demographic data (e.g., age, education, occupation) were assessed at the beginning of the marriage. The mean age of the males in our working sample was 27 years old and ranged from the late teens to the 60s. They were predominantly Caucasian—only three minorities—and the majority of their religious choices were split between "none" and Protestant. Their mean educational level achieved was "some college," ranging from the eighth grade to the doctoral level. Their mean occupational level was lower-middle class and ranged from lower class to upper-middle class (seven class levels were considered: lower, working, low to upper middle, low-upper, and upper).

The mean age of the women in our final sample was 24 and ranged from the late teens to the 60s. Their race and religious affiliations were similar to the males. Their mean educational level achieved was "some college" and ranged from the eighth grade to the master's level. Their mean occupational level was middle-middle class and ranged from working class to upper-middle class.

The average length of time for knowing each other before marriage was 2 years. This ranged from a minimum of 6 months to a maximum of 8 years. Twelve percent did not get engaged. For those couples who did get engaged, the average length of engagement time before marriage was 5 months and ranged from less than 1 month to over 3 years.

Initial Data Collection

We used the Bentler Psychological Inventory (BPI), which assesses 28 personality traits (Comrey, Backer, & Glaser, 1973). The BPI consists of 680 pairs of statements, with each pair representing two poles of a single dimension. The respondent is asked to choose the one item of each pair that most closely reflects himself/herself. Although the BPI

was developed with multivariate methods, the items are fairly face valid so that the results indicate how the person views himself/herself. Table 1 provides a listing of the 28 trait-scale names. Brief descriptive examples of high and low rankings on each scale are given in columns 2 and 3. Columns 4 and 5 show the internal consistency coefficients for a standardization sample of the self- and peer-inventories, respectively (Bentler, Note 2). The last column pro-

Table 1
Personality Trait-Scale Descriptions

Scale name	Rank		Statistical characteristics ^a		
	High	Low	Self K-R 20 ^b	Peer K-R 20 ^b	Self- Peer <i>r</i>
Agility	Quick reflexes	Clumsy	.92	.93	.50
Ambition	Seek status; want to be well-known	Satisfying position is OK	.95	.95	.47
Art Interest	Like museums and the fine arts	Not concerned with art exhibits	.94	.94	.69
Attractiveness	Nice face; sexy	Homely; plain; average features	.96	.95	.39
Body weight	Thin; skinny	Husky; fat	.98	.97	.77
Cheerfulness	Playful; happy	Moody; serious	.93	.94	.54
Clothes consciousness	Fastidious dresser; neat	Sloppy; sometimes unkempt	.95	.95	.73
Congeniality	Good-natured	Impatient; stern	.88	.92	.44
Deliberateness	Premeditated; careful	Impulsive; rash; make few plans	.93	.95	.41
Diligence	Energetic	Lazy; dawdler	.94	.94	.50
Extraversion	Pushy; talker	Shy; reserved	.95	.95	.64
Flexibility	Give in; pliable	Obstinate	.87	.92	.43
Generosity	Spender; sharer	Selfish; miserly	.93	.95	.48
Intelligence	Intellectual	Mentally average	.93	.91	.52
Invulnerability	Thick-skinned; not easily hurt	Sensitive; easily hurt	.94	.95	.38
Law Abidance	Must be legal	Rules can be bent	.89	.92	.60
Leadership	Domineering	Follower	.95	.96	.60
Liberalism	Like social protest and change	Conservative	.89	.90	.65
Masculinity	Read sports page; mannish	Like to sew; read social page	.96	.96	.94
Objectivity	Scientific	Superstitious	.87	.87	.60
Orderliness	Organized; neat	Messy; careless	.96	.97	.55
Perceptiveness	Empathetic; aware	Can't tell other's feelings	.86	.93	.25
Religious commitment	Believe in God; like church	Atheist; don't go to church	.94	.94	.76
Self-acceptance	High self-esteem; happy with self	Feel worthless; bad self-image	.95	.97	.50
Stability	Relaxed; composed	Tense; nervous	.94	.95	.44
Thriftiness	Buy on sale; save coupons	Don't shop around	.92	.93	.48
Travel interest	Take trips often; traveler	Stay at home; rarely take trips	.92	.95	.44
Trustfulness	Faith in people	Often disbelieve	.93	.93	.41

^a Based on Bentler (Note 2); *n* = 216.

^b Kuder-Richardson 20 internal consistency coefficient.

vides the peer-self correlations based on the standardization data, representing the cross-correlations between BPI and the Bentler Interactive Psychological Inventory (BIPI) scales. The BIPI is identical to the BPI but is directed at the description of another person. It is generally filled out by a peer or knowledgeable observer of a subject. In the current study, the BIPI was filled out by a friend of the husband or wife in reference to the husband or wife, respectively, representing peer evaluations of personality.

The Sexual Behavior Inventory consists of a listing of 20 sexual activities, based on Bentler (1968a, 1968b). The couple was asked to answer in a yes/no manner which of these behaviors they, as a couple, have engaged in.

Background information was obtained for both individual partners, as well as on their relationship. For each individual information was collected on age, height, weight, race, religion, education, occupation, previous divorce (or death of spouse), previous children, and parental divorces. In regard to the couple, it was ascertained how long they had known each other, been engaged, and, if applicable, lived together before marriage.

Follow-Up Data Collection

Our follow-up questionnaire had two sections. The first section consisted of the Locke and Wallace (1959) Marital Adjustment Scale, which we scored by their methods. The second section consisted of 3-point ratings (no problem = 0, moderate problem = 1, extreme problem = 2) of 19 potential problem areas. This was scored by weighted summation across all problem areas. One questionnaire was completed by each still-married couple and by as many divorced people as could be located. A score was calculated for each section on which questionnaires had been completed by both partners; in a divorced marriage, we averaged the respective scores. In total, we obtained this information on 68 couples. Although it might have been preferable to have had each partner of an intact marriage complete a questionnaire separately, we felt that we would have no control over whether in fact they had been completed independently. To determine how much variation was introduced by having still-married couples complete a single questionnaire, we included an item that asked them to assess the extent of agreement or disagreement that they had experienced while completing the form. On a 6-point rating scale (from always agreed to always disagreed), 98% reported that they had always agreed or almost always agreed. The remaining 2% indicated that they had occasionally disagreed. It can be argued that having the questionnaire completed differently by the two groups may have some effect on the validity of the measure. Although this hypothesis cannot be confirmed or denied, the procedure we used seemed the most feasible.

A factor analysis of the adjustment scale revealed substantial positive loadings of all items on the first

unrotated factor, which accounted for 75% of the variance. Thus, the scale assesses a fairly unidimensional quality. Since the first and second part scores were significantly correlated, $r(66) = -.78, p < .001$, indicating that marital adjustment and lack of problems were positively related, a composite score was obtained by averaging the normalized adjustment and lack of problems score. A probabilistic lower bound to reliability of the composite in the population (ρ) is given by $\hat{\rho}_{.99} = .78$; that is, $\rho(\hat{\rho} \leq \rho) \geq .99$ (Woodward & Bentler, in press), indicating that the composite score possesses adequate internal consistency. This composite dependent variable differentiated the married from divorced groups quite well, $t(66) = 9.54, p < .001$.

Results

Evidence for the first hypothesis comes from within-group similarities between partners, with the homogamy hypothesis predicting greater initial similarity of partners among couples longitudinally successfully married than among those who become divorced or later separate.

In regard to this hypothesis, five of the background items (age, education, occupation, previous divorce, and previous children) had significantly positive husband/wife correlations for the married couples compared with four (age, education, parental divorce, and previous children) for the divorced couples. Table 2 shows these correlations in the top portion of columns 1 and 2; a one-tailed test in the hypothesized direction was used to assess significance. The third column in this table shows the Fisher r to z conversion for test for significant differences between correlations, also using a one-tailed test. Only the age correlation for the married group was significantly more positive than it was in the divorced group. Looking at the personality trait section of columns 1 and 2 of Table 2, it can be seen that there were 10 significant husband/wife correlations for the married couples, and there were 6 for the divorced couples, again using a one-tailed test. Four traits (ambition, liberalism, religious commitment, and travel interest) had significant positive correlations for both groups. The only substantial negative correlation was on stability for the divorced group: This seems to be an area in which a mismatch contributed to marital failure. Column 3 shows that three

Table 2
Intracouple Correlations and Mean Differences

Item	Husband-wife <i>r</i>		<i>z</i> between Married & Divorced <i>r</i>	<i>M</i> _{diff} between husband and wife	
	Married (<i>n</i> = 53)	Divorced (<i>n</i> = 24)		Married (<i>df</i> = 104)	Divorced (<i>df</i> = 46)
Background					
Age	.84***	.45*	2.83**	2.76**	3.05**
Height	-.08	-.20	.47	10.98***	6.76***
Weight	.13	-.20	1.28	11.70***	7.92***
Education	.28*	.37*	-.39	.67	1.10
Occupation	.59***	.30	1.42	-2.20*	.36
Previous divorce	.64***	.60**	.25	.59	.30
Previous children	.57***	.59**	-.12	.42	1.22
Parents' divorce	.02	-.37	1.57	.50	.57
Personality traits					
Agility	-.11	.09	-.77	3.82***	3.63**
Ambition	.40***	.36*	.18	3.51**	1.92
Art interest	.53***	.02	2.19*	-2.47*	-1.95
Attractiveness	.59***	.07	2.34**	-1.80	-.99
Body weight	-.09	-.21	.47	-2.47*	-.29
Cheerfulness	-.01	.03	-.15	.95	.18
Clothes consciousness	.02	.52**	-2.14	-3.18**	-1.03
Congeniality	.24*	.04	.79	2.33*	.48
Deliberateness	-.18	-.34	.66	.32	1.14
Diligence	.10	-.02	.46	1.13	.48
Extraversion	.19	-.25	1.72*	-2.29*	-.06
Flexibility	.09	.07	.08	-1.23	-1.57
Generosity	.23*	.13	.40	-2.24*	-.62
Intelligence	.21	-.01	.86	-.17	.08
Invulnerability	.00	.22	-.86	4.36***	6.23***
Law abidance	.22	.21	.04	-2.70**	-1.65
Leadership	.14	-.06	.77	3.21**	1.94
Liberalism	.44***	.40*	.19	-.58	.14
Masculinity	-.22	-.17	-.20	27.77***	17.10***
Objectivity	.33**	.07	1.05	3.86***	3.56**
Orderliness	.07	.02	.19	-1.18	.55
Perceptiveness	-.06	-.12	.23	-2.81**	.00
Religious commitment	.48***	.69***	-1.25	-1.66	-1.03
Self-Acceptance	.00	.17	-.66	.47	1.47
Stability	-.13	-.46	1.41	3.20**	2.06*
Thriftiness	.25*	-.05	1.17	-.65	.15
Travel interest	.35**	.43*	-.36	.52	.91
Trustfulness	.10	.37*	-1.11	-2.77**	-3.15**

* $p < .05$.** $p < .01$.*** $p < .001$.

correlations (art interest, attractiveness, and extraversion) were significantly more positive, using a one-tailed test, for the married, as compared with the divorced, group. Clothes consciousness, on the other hand, seems to be substantially contrary to the homogamy hypothesis, since a larger, positive correlation

was obtained for the divorced, compared with the still-married, group. As another way of looking at this issue, a sign test on the husband/wife correlations, compared across the married and divorced groups, showed that the still marrieds had more positive correlations than the divorced ($p < .05$, one-tailed test).

Turning next to mean differences rather than correlational similarities, the upper portion of column 4 in Table 2 shows that there were four background areas that were significantly different between husbands and wives for the married group; these differences were on age, height, weight, and occupation. Two-tailed *t* tests were used on these mean comparisons. A positive sign indicates the husband had the higher score, and a negative sign means that the wife had the higher score. Column 5 shows that for the divorced group three areas were significantly different; these were age, height, and weight. Looking

next at personality traits, column 4 shows that there were 16 significant mean differences between husbands and wives among the still-married group, whereas there were 6 significant mean differences for the divorced group, as can be seen in column 5. The groups were significantly different in regard to the number of trait differences when the results were considered statistically, $\chi^2(1) = 6.06$, $p < .025$.

We then looked at whether the magnitudes of the husband/wife discrepancies were statistically different between the married and divorced groups. Although none of the dis-

Table 3
Peer and Self-Trait-Rating Comparisons

Trait	<i>r</i> between self and peer				<i>t</i> between married and divorced		The average of male and female (Self-Peer) ^a (<i>df</i> = 18)
	Male		Female		Self-Peer ^a		
	Married (<i>n</i> = 14)	Divorced (<i>n</i> = 6)	Married (<i>n</i> = 15)	Divorced (<i>n</i> = 9)	Male (<i>df</i> = 18)	Female (<i>df</i> = 22)	
Agility	.43	.44	.65**	.62*	-.09	.74	.44
Ambition	.44	.86*	.47*	.51	.70	-1.14	.18
Art interest	.71**	.71	.82***	.27	-.50	-.61	-.94
Attractiveness	.23	.81*	.36	.83**	.70	-.03	.58
Body weight	.53*	.92**	.88***	.86***	.24	-.10	-.88
Cheerfulness	.41	.23	.06	-.03	-2.69**	-.07	-2.04*
Clothes consciousness	.41	.82*	.69**	.88***	1.55	.57	1.18
Congeniality	.10	.79	.09	.73*	1.77	1.33	1.23
Deliberateness	.58	-.09	.30	.70*	-.73	-.23	-.79
Diligence	.72**	.87*	.54*	.74*	.18	.47	.49
Extraversion	.40	.93**	.66**	.67*	1.86	-.33	.44
Flexibility	-.05	.78*	.10	.37	2.60	.62	2.15
Generosity	.62**	.88**	.64**	.73*	1.64	.06	.87
Intelligence	.56*	.69	.72***	.31	.94	.13	.74
Invulnerability	-.19	-.49	.69**	.64*	-.37	-.55	-.66
Law abidance	.67**	.04	.60**	.88***	-2.07*	.95	-1.52
Leadership	.24	-.16	.48*	.57	-.76	1.01	-.30
Liberality	.66**	.20	.80***	.89***	-1.34	1.29	-.74
Masculinity	.41	.04	.77***	.44	-1.51	-1.87*	-2.11*
Objectivity	.13	.73*	.41	.79**	1.05	.55	1.30
Orderliness	.46*	.67	.83***	.85**	.56	.43	-.35
Perceptiveness	.22	-.36	.19	.00	1.75	.80	1.20
Religious commitment	.68**	.45	.73***	.74*	.15	-.22	-.06
Self-Acceptance	.66**	.25	.60**	.60*	-1.67	-1.35	-1.71
Stability	.10	.82*	.45*	.71	.80	.73	1.21
Thriftiness	.55*	-.48	.44*	.51	-1.10	.11	-1.00
Travel interest	.17	.00	.42	.53	-.17	.69	.07
Trustfulness	.39	.17	.33	.55	-.39	-.62	-.78

Note. All significance tests are one-tailed in the direction of our hypothesis.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

crepancies were found to be significantly different, taken trait by trait, a sign test revealed that the husband/wife differences were significantly smaller overall for the divorced compared to the married group ($p < .05$, two-tailed). Combined with the correlational results, it seems that mean differentiation in personality between husbands and wives who are successfully married coexists with relative agreement correlationally, in self-perception.

Using tests appropriate for nominal data, the background items "race" and "religion" showed no significant husband/wife variation or difference between married and divorced groups.

Our second hypothesis, concerning peer/self congruity as a measure of accuracy of self-perception, is mildly supported by our findings. Referring to Table 3, columns 1 and 2 for males and 3 and 4 for females show that although there were more significantly non-zero peer/self correlations among both males and females of the married group than the divorced group, the difference in number was not significant for either males or females. Using again the Fisher r to z conversion, we find that thriftiness, for the males, and art interest, for the females, showed a significantly higher peer/self correlation within the married group as compared to the divorced group. (Although not related to marital differences, women showed a significantly larger number of nonzero peer/self correlations than did the males; apparently, there is greater peer/self agreement regarding the wife's personality than the husband's personality.) Turning to the mean scores, it was noted that the married males showed a significantly lower peer/self discrepancy (squared) on cheerfulness and law abidance, as compared to the divorced males, as can be seen in column 5 of Table 3. One trait, masculinity, showed analogous results for the females (column 6). Flexibility, for the males, turned out to be substantially contrary to our hypothesis. When the peer/self discrepancies (squared) on a given trait were averaged for the couple, cheerfulness and masculinity were significantly less discrepant in the married than in the divorced group (see Table 3, column 7). Flexibility, on the other hand, showed a sub-

stantially greater discrepancy for the married as compared with the divorced group, which is contrary to our hypothesis.

Our third hypothesis, concerning living together before marriage, was not supported by the data. For married couples, 24 lived together before marriage and 29 did not. For Divorced couples, 14 lived together before marriage and 10 did not, $\chi^2(1) = .66$, ns. In regard to other couple background characteristics, we found no significant differences between the divorced and still-married groups for the length of time they had known each other, been engaged, or lived together before marriage (if in fact they did). Although none of these mean differences was significant, the still-married couples tended to have known each other longer and to have been engaged longer than the divorced group. The opposite tendency was true for the duration of living together before marriage (if they did). The divorced group had lived together longer than the still marrieds. One other background question did turn out to be a statistically reliable predictor of marital success. The couples who have remained together more often had one or both partners previously widowed than did the couples who had divorced. In regard to the Sexual Behavior Inventory, there was no significant differentiation revealed between the married and divorced groups on this measure.

Our fourth hypothesis, concerning types of problems faced by divorced and married couples, shows some clear distinctions. Table 4 shows the mean values of the married and divorced groups for each problem area (in columns 1 and 2, respectively). Column 3 shows the t values for each problem area between the two groups. There were six areas—venereal disease, desire for children, drunkenness, gambling, male sent to jail, and in-laws—which were essentially no more of a problem for the divorced group than the married group. Twelve areas proved to be significantly greater problems for the divorced, compared to the married, group. These could be roughly considered to be interpersonal volitional areas in which conflict or disturbance can occur, for example, sex relations, lack of mutual affection, bickering, selfishness, and desire for independence. One area—

Table 4
Problem Ratings at Follow-Up

Problem	<i>M</i> problem score ^a		<i>t</i> between married and divorced (<i>df</i> = 66)
	Married (<i>n</i> = 53)	Divorced (<i>n</i> = 15)	
Attention to another	.2	.9	-4.17***
Mutual affection	.1	1.1	-5.13***
Adultery	.1	.9	-5.03***
Sex relations	.4	.9	-2.73**
Venereal disease	.0	.0	.00
Desire for child	.2	.3	-.44
Finances	.3	1.0	-4.34***
Nonsupport	.0	.5	-3.80***
Drunkenness	.2	.3	-.54
Drug abuse	.0	.2	-2.65**
Gambling	.0	.1	-.70
Sent to jail	.0	.1	-.69
Friends	.1	.7	-4.92***
Selfishness	.3	1.0	-4.37***
In-laws	.3	0.5	-.73
Ill health	.2	.0	+2.41*
Bickering	.3	.6	-2.04*
Independence	.3	1.0	-3.87***
Career conflicts	.1	.5	-2.72**
Other	.3	.5	-.84
Problem score	3.5	11.3	-6.00***

Note. All significance tests are two-tailed.

^a 0 = no problem; 1 = moderate problem; 2 = extreme problem.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

ill health—was found to be a significantly greater problem for the married than for the divorced groups. This is more of an involuntary problem and seems to have drawn the couples together.

The successful longitudinal prediction of marital success requires significant differentiation of means when comparing still-married and divorced groups. Columns 1 and 2 of Table 5 show the *t* values for these mean comparisons for the males and females, respectively. A positive sign indicates that the married group had the larger value, whereas a negative sign means the divorced group had the greater value. We used one-tailed tests of significance on the background items, since we had a priori expectations from previous research (e.g., Murstein & Glaudin, 1966), which we will address more fully in the Discussion section. The only significant difference on background item means between married and divorced males was on parental divorce.

The married males had fewer parental divorces than the divorced males. For females there were five significant mean differences. The married females were older, had higher educational and occupational levels, had more previous children, and had fewer parental divorces than the divorced females. Comparing mean scores on personality traits across the two groups, using a two-tailed test of significance, the following picture emerged: The married males showed significantly less extraversion, invulnerability, and orderliness compared with the divorced males. The married females showed significantly greater clothes consciousness and congeniality compared to the divorced females.

Next, we evaluated the longitudinal predictability of marital success with initial personality variables obtained at the time of marriage, variable by variable. The correlations between the composite dependent marital adjustment score and prior personality

Table 5
Predictive Correlations and Mean Differences Between Married and Divorced Groups

Item	M_{diff} between married and divorced		r with the composite adjustment score		z between male and female r
	Males ($n = 53$)	Females ($n = 24$)	Males ($n = 68$)	Females ($n = 68$)	
Background					
Age	1.36	2.11*	.29**	.27**	.12
Height	-.07	1.17	-.12	.10	-1.26
Weight	-1.60	.23	-.05	.04	-.51
Education	.91	1.77*	.09	.00	.51
Occupation	-.70	1.76*	.02	-.10	.69
Previous divorce	.52	.41	.13	.13	.00
Previous children	1.16	2.55**	.21	.33**	-.74
Parents divorced	-2.30*	-2.00*	-.21	-.01	-1.16
Personality traits					
Agility	-1.02	.02	-.04	-.03	-.06
Ambition	-1.16	-1.93	-.19	-.24*	.30
Art interest	-.35	-1.07	-.10	-.28*	1.07
Attractiveness	-.31	-.62	-.12	-.14	.12
Body weight	-1.37	.22	-.01	-.02	.06
Cheerfulness	.24	1.27	-.01	.14	-.86
Clothes consciousness	.28	2.05*	-.01	.27*	-1.64
Congeniality	-.10	2.52*	.06	.14	-.46
Deliberateness	.38	1.60	.25*	.04	1.23
Diligence	-.28	-.53	.02	.17	-.86
Extraversion	-2.20*	-.59	-.29**	.05	-1.99*
Flexibility	.56	-.15	.02	-.06	.46
Generosity	-.02	.86	-.07	.18	-1.44
Intelligence	-1.06	-.92	.01	-.23*	1.39
Invulnerability	-2.05*	.19	-.17	.19	-2.08*
Law Abidance	.77	.87	.11	.15	-.23
Leadership	-.81	-.69	-.16	.00	-.92
Liberalism	-1.74	-1.37	.00	-.14	.80
Masculinity	.84	-.33	-.03	.09	-.69
Objectivity	-.04	.95	.12	.23*	-.65
Orderliness	-2.18*	-.62	-.12	.05	-1.03
Perceptiveness	-.86	1.79	-.06	.16	-1.26
Religious commitment	.64	.91	.01	.04	-.17
Self-acceptance	-.07	1.25	.02	.14	-.69
Stability	.78	1.19	.13	.24*	-.59
Thriftiness	-1.23	-.97	.14	.05	.52
Travel interest	-.98	-.68	-.12	-.02	-.57
Trustfulness	1.82	.71	.11	.00	.63

* $p < .05$.

** $p < .01$.

scores were obtained separately for males and females (see columns 3 and 4 of Table 5). A two-tailed significance test was used. Among males, two significant correlations were found on extraversion and deliberateness. The more adjusted and happy the marriage, the more deliberate and less extraverted were the males. Among females, six significant correlations

were found for ambition, art interest, clothes consciousness, intelligence, objectivity, and stability. The higher the marital adjustment score, the more the females reported themselves to be clothes conscious, objective, and stable while reporting relatively less ambition, art interest, and intelligence. Again using the r to z conversion, column 5 shows the z

Table 6
*Stepwise Selection Regression Equation: Using
 Initial Pool of All Variables*

Variable	Beta	Sign
Previous children—female	.21*	1
Deliberateness—male	.23*	1
Ambition—female	-.19	-1
Objectivity—female	.33**	1
Clothes consciousness—female	.45***	1
Orderliness—male	-.37***	-1
Masculinity—female	.33**	1
Intelligence—female	-.26*	-1
Thriftiness—male	.24*	1
Flexibility—male	-.23*	-1

Source	SS	df	MS	F
Signed, equal weight	25,887.3	1	25,887.3	60.31***
Differential weights	1,829.4	9	203.3	.47
Total				
regression	27,716.7	10	2,771.7	6.45***
Residual	23,607.2	55	429.2	

* $p < .05$.

** $p < .01$.

*** $p < .001$.

values when comparing the correlations of the males and females. Two correlations were significantly different between the males and the females. The males had negative correlations on extraversion and invulnerability, whereas the females had significantly different, positive correlations. The greater number of significant correlations on personality traits among females reflects the apparent greater predictability of marital success among women than men.

When the composite dependent variable was correlated with the background variables, by sex, there were two significant correlations for females and one for males. For the males, being older is predictive of marital adjustment, whereas for females, being older and having previous children are predictive of marital adjustment. There were no significant differences between males and females for any of the background correlations, using the r to z conversion.

Several longitudinal prediction equations involving linear combinations of predictor variables have been generated from our data, using as dependent variables (a) the com-

posite adjustment score and (b) being married versus divorced at follow-up. The most concise summarization of these results can be stated for the prediction equations using background and personality data from all male and female subjects, focusing on the continuous dependent variable (the composite score). A stepwise-regression procedure was used, which showed significant increments in prediction up to the point at which the procedure was terminated, due to a conservative judgment regarding stability of results on potential cross-validation. With 10 predictor variables the multiple correlation was .74, $F(10, 55) = 6.45$, $p < .001$. The beta weights for this equation are shown in the second column of Table 6, and the F test associated with this multiple regression model is given in the third row of the analysis of variance breakdown. All but one of the predictors contributed significantly ($p < .05$) to the multiple regression equation, whereas the weight of the remaining variable was 1.85-times its standard error.

To assess in finer detail the nature of the regression, we used a new regression model developed by Bentler and Woodward (Note 3; Woodward & Bentler, Note 4). This model can be written in the form $y = \beta^*Xt + X(\beta - \beta^*t) + e$, where y is the standardized dependent variable, X is the matrix of standardized predictor variables, and e is the residuals. The vector of the beta weights is the usual optimal vector of least-squares coefficients. The vector t is a sign vector with elements ± 1 , chosen to optimize prediction of y from X , and β^* is a weight associated with this regression. The Bentler-Woodward regression method in essence partitions the usual sum of squares due to regression into two additive components: one, associated with β^*Xt , and the second, associated with $X(\beta - \beta^*t)$. The first component is determined such that the sum of squares associated with that component is maximized; it represents the optimal prediction possible with weights β^*_i that are equal for all i predictors, except for sign. The second component assesses whether differential weighting contributes a significant increment to prediction beyond that possible with the signed, equal weights.

The Bentler-Woodward regression solu-

tion of Table 6 was obtained after the stepwise regression method yielded the 10 predictors shown in the table. It determined that the optimal sign weights for the regression were those given in the right-hand part of the table. In this case these signs are identical to those of the beta weights; this is not a necessary feature of the approach. The analysis of variance at the bottom of Table 6 shows that the null hypothesis that the contribution of the signed, equal weights to prediction is zero could be rejected at a high level of statistical significance ($p < .001$). On the other hand, the analysis of variance revealed that the contribution of variance revealed that the contribution of differential weighting, over and above the contribution of the signed, equal weights, was not statistically reliable. Thus, although the ordinary multiple regression procedure yielded a significant regression of criterion on predictor variables, the more fine-grained analysis reveals that virtually the only source of such prediction is associated with the signed, equal weights.

The proportions of variance associated with the Table 6 results mirror the statistical results. The correlation of the signed equal-weight composite with the criterion was .71, yielding an $R^2_{\beta^*} = .50$. The contribution of differential weighting was only $R^2_{\beta} - R^2_{\beta^*} = .04$. Of course, these two components of regression add to the total variance accounted for by the least-squares beta weights $R^2_{\beta} = .54$. As a consequence of these results, one would not expect the differential weights to cross-validate. Another way to assess the potential shrinkage due to using the highly tuned beta weights is to evaluate the adjusted R^2_{β} , which represents an unbiased estimator of the population squared multiple correlation (Olkin & Pratt, 1958). It is an adjustment of the R^2_{β} value using the number of independent variables and the sample size. Specifically, the equation we used was

$$\text{adjusted } R^2_{\beta} = 1 - \left(\frac{N-3}{N-k-1} \right) (1 - R^2_{\beta}) \\ - \left(\frac{N-3}{N-k-1} \right) \left(\frac{2}{N-k+1} \right) (1 - R^2_{\beta})^2,$$

where k equals the number of predictor variables in the equation and N equals the total

number of cases. Our adjusted $R^2_{\beta} = .46$, a value in the same range as $R^2_{\beta^*}$.

Turning now to the substantive meaning of the regression described in Table 6, the variables included in the regression were those obtained from women (presence of previous children, less ambition, objectivity, clothes consciousness, masculinity, and less intelligence) as well as from men (deliberateness, less orderliness, thriftiness, and less flexibility). In spite of the variable-by-variable results mentioned previously, the simultaneous approach indicates that self-reported personality variables contributed the major share to the longitudinal predictability of marital success, with background variables being barely represented in the multiple regression. (Only presence of previous children, among females, contributed significantly.) Furthermore, the majority of significant predictor variables came from the women's data, suggesting that variation in longitudinal marital success is a greater consequence of women's attributes than that of men. This effect was more dramatically observed in stepwise regression equations based on larger numbers of variables. For example, the 15-variable equation ($R = .80$) consisted of 10 variables from women and only 5 from men. Although the dimension of traditionality in sex role orientation (Ellis & Bentler, 1973) may be relevant to understanding these results (see Discussion), the differential prediction effect does not seem explainable.

One goal in constructing optimally weighted prediction equations is to find one that maximizes the amount of variance (R^2_{β}) accounted for by a small number of variables. Statistically, R^2_{β} can be decomposed into the sum of the products of the simple correlation coefficients and the beta weights for each variable in the equation. We arbitrarily set a minimum level of .03 for this product and located 15 variables from an equation based on the best 30 variables that met this criterion. These variables thus showed a minimally adequate beta weight in the set of all variables as well as a minimally adequate zero-order criterion correlation. Using these 15 variables in a stepwise regression manner led to the equation of Table 7. Only 12 variables became part of the final equation, since

Table 7

Stepwise Selection Regression Equation From the Pool of 15 Variables that Maximized the Beta X Simple Correlation

Variable	Beta	Sign
Previous children—female	.24*	1
Objectivity—female	.27*	1
Clothes consciousness—female	.32**	1
Extraversion—male	-.10	-1
Masculinity—female	.40***	1
Intelligence—female	-.39***	-1
Parents divorced—male	-.15	-1
Cheerfulness—female	.20*	1
Orderliness—male	-.29*	-1
Thriftness—male	.21*	1
Invulnerability—male	-.19	-1
Perceptiveness—female	.16	1

Source	SS	df	MS	F
Signed, equal weight	24,655.5	1	24,655.5	58.70***
Differential weights	4,407.8	11	400.7	.95
Total regression	29,063.3	12	2,421.9	5.77***
Residual	22,260.6	53	420.0	

* $p < .05$.** $p < .01$.*** $p < .001$.

3 of the 15 did not meet our minimum entry criterion in the stepwise inclusion ($F > 1.0$). Table 7 exhibits virtually the same pattern of results regarding the relative merits of various sets of weights as does Table 6. In this case,

$$R^2_{\beta^*} = .48, \quad R^2_{\beta} - R^2_{\beta^*} = .09, \quad R^2_{\beta} = .57, \\ \text{adjusted } R^2_{\beta} = .48.$$

Again, the signed, equal weights carry the lion's share of the total prediction, as well as the only statistically reliable effect.

To evaluate whether the stepwise regression procedure was somehow unfairly selecting from the pool of background or personality variables, or from variables associated with a given sex, we ran separate stepwise regression equations on the background variables and traits for the males and females. From each of these four equations, we selected the 2 best predicting variables according to an F test on the beta weights. Combining these 8 variables with the 10 in the first equation (Table 6) and the 12 in the second equation (Table 7), we arrived at a pool of 21 variables.

This pool was considered to adequately represent the background variables. A stepwise regression was run on this set of variables. The results are shown in Table 8. This equation uses 11 variables and was chosen since it maximized the overall F value associated with R^2_{β} . Perusing the substantive content, the sex difference is noted again, with the males being represented by 3 of the 11 variables and the females representing the majority by 8 of the variables. This result is not due to an initial bias in the pool of 21 variables, since they were fairly evenly divided by sex (10 males and 11 females). Turning finally to the quantitative results, we find that the analysis of variance mirrors the results of Tables 6 and 7. We have

$$R^2_{\beta^*} = .49, \quad R^2_{\beta} - R^2_{\beta^*} = .08, \quad R^2_{\beta} = .57, \\ \text{and adjusted } R^2_{\beta} = .49.$$

(It should be noted that the total degrees of freedom associated with Table 8 is two greater than in either Table 6 or Table 7. This is due to not having a complete data set on every

Table 8

Stepwise Selection Regression Equation of the Combined Pool of 21 Variables

Variable	Beta	Sign
Previous children—female	.33**	1
Ambition—female	-.21*	-1
Objectivity—female	.29**	1
Thriftness—male	.28**	1
Perceptiveness—female	.20	1
Intelligence—female	-.33**	-1
Masculinity—female	.45***	1
Cheerfulness—female	.21*	1
Parents divorced—male	-.15	-1
Clothes consciousness—female	.26*	1
Orderliness—male	-.21*	-1

Source	SS	df	MS	F
Signed, equal weight	25,818.1	1	25,818.1	63.75***
Differential weights	3,942.9	10	394.3	.97
Total regression	29,761.0	11	2,705.5	6.68***
Residual	22,681.1	56	405.0	

* $p < .05$.** $p < .01$.*** $p < .001$.

couple and does not reflect a numerical error.)

Although we did generate several equations using our dichotomous dependent variable (still married vs. divorced), we have chosen not to report these for several reasons. Most important was the consideration that the statistical assumptions underlying this type of dichotomous regression are less sound and proven than for a continuous dependent variable. This uncertainty and lack of soundness could create difficulty in cross-validating any dichotomously based equations. For example, a simple shift in the proportion of divorced couples in a given study would modify the optimal beta weights and the overall R^2 , due to the sensitivity of the regression to the mean of the dependent variable. This is why we have chosen to report only equations based on the continuous composite dependent variable.

Discussion

The present investigation has shown some clear longitudinal evidence for the homogamy hypothesis. Our results indicate that homogamy of personality traits between marital partners, assessed at the beginning of their marriage, is evidenced to a greater degree in marriages that turn out successfully than for marriages that terminate in separation or divorce. In other words, correlational similarity between marital partners, based on personality traits measured at the beginning of a marriage, was substantially higher for couples who remained together after 4 years than couples who decided to end their marriage within that period of time. This pattern was also found for background or demographic variables. Other researchers have found similar results cross-sectionally on personality traits (e.g., Cattell & Nesselroade, 1967; DeYoung & Fleischer, 1976; Pickford et al., 1966b). Using the Guilford-Zimmerman Temperament Survey, Pickford et al. (1966b) found four significantly positive husband/wife correlations on general activity, restraint, friendliness, and personal relations for their happily married group. Our significant finding on congeniality seems similar to theirs on friendliness for happily or still-married cou-

ples. They found no significantly positive correlations, but they did report one significantly negative correlation on emotional stability for couples on the verge of separation; our divorced sample had a similarly large negative correlation on stability. Cattell and Nesselroade (1967), using the 16 Personality Factor Questionnaire (16 PF), found four factors (Affectothymia, Surgency, Protension, and Self-Sufficiency) that were significantly more positively correlated between husbands and wives in stable marriages compared to couples in unstable marriages. Both Affectothymia and Surgency seem similar to our significantly different correlations on extraversion. DeYoung and Fleischer (1976) found all positive husband/wife correlations, using the 16 PF, with 10 (Affectothymia, Intelligence, Dominance, Surgency, Super Ego Strength, Autia, Radicalism, Self-Sufficiency, Self-Concept Control, and Ergic Tension) significantly so. They found these results in their one sample of couples who were still married and did not contrast them with a group experiencing marital disharmony or divorce.

This type of analysis does not directly address the need complementarity hypothesis (e.g., Meyer & Pepper, 1977; Murstein, 1961, 1967; Rosow, 1957; Winch, 1967), since the validity of inferring needs from personality traits is not adequately known. Nonetheless, one can consider many of our trait scales to be bipolar (e.g., extraversion: high score means very extraverted, a low score means very introverted), so that evidence for a complementary trait hypothesis can be examined. For the successfully married couples, no trait was found to have a significantly negative correlation between partners. If complementarity were an important influence for a successful marriage, some significant negative correlations should have been found in this group. The divorced group, on the other hand, had one substantial negative correlation on stability. Although we do not claim that these results disprove a complementary trait hypothesis, certainly there is no evidence favoring the concept. Yet, when we look at our simple prediction correlations by sex (traits predicting the composite dependent variable), we find two traits—

extraversion and invulnerability—that have significantly different effects for the males and the females. The happy and/or maritally adjusted males were relatively introverted and vulnerable, whereas the successfully married females were relatively extraverted and invulnerable. When comparing these two types of analyses—husband/wife correlations and simple predictive correlations between individual traits and the composite dependent measure—no clear support or refutation of the complementarity hypothesis seems evident. Thus, although we have found longitudinal support for the homogamy or correlational similarity hypothesis of spouse personality traits, we can say little regarding the complementarity (of traits) notion.

Some researchers (e.g., Singh et al., 1976) have found that significant mean differentiation of marital partners on personality traits occurs more often in successfully married couples than in those who divorce. These researchers implied that this finding is incompatible with correlational similarity between successfully married spouses. Our results do not provide longitudinal support for this cross-sectional finding. We found both significant mean differentiation as well as correlational similarity on personality traits between partners within the still-married group. In addition, we found greater mean differences among the successfully married than the divorced. In other words, correlational similarity does not necessarily imply equality of trait levels; it only indicates that the traits are related in a linear manner. Apparently, correlational similarity and mean differentiation between marital partners are both conducive to subsequent marital happiness and success. How this finding is reflected in spouses' behavior toward each other and how it might contribute to marital satisfaction is unclear. Further theory development and research needs to be done to integrate and thoroughly understand this finding.

Our hypothesis concerning the effect of self-perception accuracy, as revealed by peer/self rating comparisons, was only mildly supported by our data. We speculated that the more similar peer- and self-ratings were before marriage, the more accurately that spouse would self-perceive in the marriage.

He or she should then face a marriage relationship more adaptively, resulting in greater adjustment, happiness, and ultimately success in marriage. Weigel, Weigel, and Richardson (1973), addressing somewhat similar issues, found no significant results, whereas Murstein and Beck (1972) found some general support for this idea. Although our findings were not strong, many were in the direction hypothesized, and some significantly so. Some results found contrary to this hypothesis have been previously noted. Perhaps if the proportion of our sample that did have peer assessments of their personalities had been larger, more definitive results might have been found. Of course, it is also possible that accuracy in interpersonal and self-perception may generate a more rapid separation if the marriage is in actual trouble, or perhaps the effect of accuracy depends on the particular traits involved. We have not as yet developed an adequate theory in this regard.

Our hypothesis concerning living together before marriage was not supported by the data. This is a somewhat discouraging finding in light of the currently growing movement toward trial or premarriages. The effect of living together would seem to merit further study beyond our reported results to determine if there are any specific variables in the couple that are important to making this experience a helpful or hindering influence on any potential subsequent marriage.

An analysis of problems faced by married couples compared to divorced couples revealed some clear distinctions. "Ill health" was the only significantly greater problem for the still marrieds and seemed to be an involuntary problem that helped draw the couple together. Several areas represented a significantly greater problem for the divorced group and could be roughly classified as volitional/relational problems. It is interesting to note that although "drug abuse" was a significantly greater problem for the divorced group, "drunkenness" was in absolute terms a greater problem for both groups than drug abuse. Thus, even though drug abuse has recently received much public attention, alcohol appears to be the greater problem during the first few years of marriage for both married and divorced couples.

In regard to the longitudinal predictability of marital success, we found seven background or trait variables that significantly differentiated the married females from the divorced females. Four background or trait variables significantly differentiated married from divorced males. Several of the background variable differences have been found by other investigators. The age relationship was found by the U.S. Bureau of the Census (1973), Luckey (1966), and Landis (1956, 1963). The educational difference was also noted by Murstein and Glaudin (1966), whereas Landis (1956, 1963) found results opposite to ours on education and occupational levels, probably because differing methods were used to define experimental groups. Several other studies (e.g., Pope & Mueller, 1976; Renne, 1971) have reported results similar to ours on parents' divorce. On the basis of previous research, we had expected that the married group would have had fewer previous divorces than the divorced group, but this effect was not found in our data. It indicates, somewhat happily, that previously divorced individuals are not at a special risk in future marriage.

We found three significant mean differences on traits between the married and divorced groups for the males and two for the females. Previous cross-sectional research (e.g., Cattell & Nesselroade, 1967; Pickford et al., 1966a; Singh et al., 1976) has found personality differences between married and divorced groups. Using the Guilford-Zimmerman Temperament Survey, Pickford et al. (1966a) found one significant difference between happily married males and males who were in marriages that were on the verge of divorce. This trait was personal relations, which may be related to our finding that males differed on extraversion. No significant differences were found between the groups for females in the Pickford et al. (1966a) study. Cattell and Nesselroade (1967), using the 16 PF, found five traits (Intelligence, Dominance, Protension, Shrewdness, and Self-Concept Control) that significantly differentiated stable from unstable marriages on husband/wife averages. Also using the 16 PF, Singh et al. (1976) found three traits (Premsia, Timidity, and Ergic Tension) that signifi-

cantly differentiated males in happy marriages from males in marital disharmony. Our significant differences for males on extraversion and invulnerability seem to coincide with their findings on Timidity and Premsia, respectively. These researchers found no significant differences between their sample groups for the females. We were not able to replicate the finding of Reevey (1963) regarding greater previous sexual experience in persons with an unfavorable marital happiness prediction; such sexual experience was unrelated to actual marital success or failure in our sample.

There were three significant simple predictive correlations between the composite dependent score and background or trait variables for the males. In contrast, there were eight significant correlations for the females. It could be argued that rather than reveal a linear trend, trait levels could relate to marital adjustment in some curvilinear fashion. For example, perhaps intermediate levels of a particular trait are more conducive to marital happiness than extremes in either direction. Although this is a possibility, we have chosen to look for linear relationships, since experience has shown little empirical support for curvilinear relationships on cross-validation (e.g., Wiggins, 1973).

For both mean differentiation between married and divorced groups as well as simple predictive correlations, the females showed the largest number of statistically significant results, indicating that the woman has greater predictive influence on the outcome of a marriage than does the man in the marital dyad. This sex difference was also found longitudinally for dating couples by Hill et al. (1976) and cross-sectionally by Murstein and Glaudin (1966). In our study, these trait differences between groups cannot be attributed to interactions between spouses. Rather, they precede marital interaction and apparently have some bearing on the subsequent outcome and quality of a marriage.

In her review of the marital adjustment literature, Laws (1971) showed that women adjust better to their husbands, given the traditional definition of roles in marital relationships. Similarly, in our study, women who rated their own ambition, intelligence, and

interest in art as being low also tended to be more satisfied 4 years later, and women who had children from a previous marriage also had greater solace in their new marriage at follow-up. Certain personality traits and some background items for women, then, seem to be quite consistent with a traditional definition of marital adjustment in society as it is today. Cross-validation in a population or society that has different values or role definitions may be less than fully confirmatory. For instance, if traditional female roles, the value of heterosexual relationships, and definitions of marital happiness change or become altered in some manner—as the feminist movement might desire—our findings on specific variables may not be replicated within this new cultural framework. It remains possible, however, that our general findings concerning homogamy, self-perception, problem areas, and so forth, will not be severely affected by a modification in society. Furthermore, certain results (such as on extraversion and invulnerability, which showed significantly different associations with marital adjustment by sex) indicate that nontraditionality may yield greater marital adjustment. Thus, the concept of sex role traditionality (e.g., Ellis & Bentler, 1973) cannot explain all the major results.

In choosing to look at personality trait variables for longitudinal prediction of the quality of marital success or failure, we had made the unproven—albeit reasonable—assumption that these types of variables will be more powerful predictors than simple background or demographic characteristics. To test this idea, we chose the two best predicting traits for the males and then the females. We then regressed these four traits on the composite dependent variable. This equation had an R of .54, $F(4, 63) = 6.58$, $p < .001$. We then chose the two best predicting background variables for the males and then the females. We entered these four variables into a prediction equation and obtained a multiple correlation of .42, $F(4, 63) = 3.39$, $p < .025$. The first equation accounted for 29% of the criterion variance, and the latter equation accounted for 18% of the variance. It seems clear when comparing these two equations that the traits have a much greater longi-

tudinal predictive effect than do the background/demographic variables.

Previous prediction equation research (Karlsson, 1951; Kelly, 1939; Locke & Karlsson, 1952; Terman & Oden, 1947) have found multiple correlations ranging from .32 to .59 using personality and/or background variables. Our equations seem to be a fair improvement over these, with multiple correlations in the .75 range. Unfortunately, our sample size was too small to cross-validate these equations, and thus this task must be left for later research. It should be noted that as a rule of thumb, for beta weights to be reliable estimates of the population, there should be about 30 cases for each predictor variable. Since our sample size does not permit this, our betas are liable to fluctuate to some extent on replication. To assess whether the optimal beta weights indeed represent some important feature of the data rather than accidental noise, we calculated the unbiased estimator of the population squared multiple correlation. This indicated that about 8%–10% of the predictive variance was due to overfitting in the sample. In view of our use of stepwise regression, which also capitalizes on chance to some extent, we would expect the population R of about .69 to be an upper-bound value rather than a true unbiased estimator that equally overshoots and undershoots the true value in its estimates. A completely different coefficient is given by the sample cross-validity correlation, which estimates the effect of using the sample regression equation in the population. Using the conservative Darlington (1968) coefficient, we obtained predicted correlations of .59, .58, and .61, somewhat below the unbiased estimates for Tables 6, 7, and 8 but still substantial in magnitude. We also used a novel regression method from Bentler and Woodward (Note 3), which determines the optimal prediction possible by using signed, equal weights rather than beta weights. This procedure found that the only significant contribution to prediction was being made by the signed, equal weights, and that any further contribution due to differential weighting simply represented statistical noise. Consequently, we would urge the use of signed weights rather than beta weights in future

replications, since these signs appear to carry all the reliable predictive information. With such weights, one obtains $R \cong .70$, still a substantial improvement over previous research. Of course, replication of the signs in new samples is a clear task for future research.

Our emphasis on predicting marital success or failure from factors assessed at the beginning of a marriage presupposes that marriage may somehow alter the personality traits of each partner. Further study needs to be done regarding how such traits are changed, if in fact they are. Does a successful marriage alter one's personality traits in a given direction—for example, in a "regression effect" toward the mean of both partners—and does the change depend on the success of the marriage? Are traits affected differentially by various marital experiences?

The interaction of two people in that uniquely intimate and intensive situation called marriage is an immensely complicated process that draws on many factors of each member's background and personality, as well as more situational characteristics. A marriage is the result of what each partner brings to the marriage and what they make of it, once together. We have focused on what each partner has brought to the marriage in order to understand the variations that seem to limit the best that individuals can subsequently make of their marriage.

Reference Notes

1. Holz, R. F. *Homogamy and heterogamy in the marital dyad: The effects of role on need disposition*. Paper presented at the meeting of the American Sociological Association, San Francisco, 1969.
2. Bentler, P. M. *On the multimethod existence of 29 personality dimensions: Data for the situation versus trait controversy*. Invited research address presented at the meeting of the Western Psychological Association, Portland, Oregon, April 1972.
3. Bentler, P. M., & Woodward, J. A. *Regression on linear composites: Statistical theory and application*. Manuscript submitted for publication, 1977.
4. Woodward, J. A., & Bentler, P. M. *Unit-weight composites with maximum variance*. Paper presented at the meeting of the Psychometric Society, Chapel Hill, North Carolina, June 1977.

References

Barton, K., & Cattell, R. B. Real and perceived similarities in personality between spouses: Test

- of "likeness" versus "completeness" theories. *Psychological Reports*, 1972, 31, 15-18.
- Bentler, P. M. Heterosexual behavior assessment—I. Males. *Behaviour Research and Therapy*, 1968, 6, 21-25. (a)
- Bentler, P. M. Heterosexual behavior assessment—II. Females. *Behaviour Research and Therapy*, 1968, 6, 27-30. (b)
- Burgess, E. W., & Wallin, P. *Engagement and marriage*. Philadelphia: Lippincott, 1953.
- Cattell, R. B., & Nesselroade, J. B. Likeness and completeness theories examined by sixteen personality factor measures on stably and unstably married couples. *Journal of Personality and Social Psychology*, 1967, 7, 351-361.
- Comrey, A. L., Backer, T. E., & Glaser, E. M. *A sourcebook for mental health measures*. Los Angeles: Human Interaction Research Institute, 1973.
- Darlington, R. B. Multiple regression in psychological research and practice. *Psychological Bulletin*, 1968, 69, 161-182.
- DeYoung, G. E., & Fleischer, B. Motivational and personality trait relationships in mate selection. *Behavior Genetics*, 1976, 6, 1-6.
- Ellis, L., & Bentler, P. M. Traditional sex-determined role standards and sex-stereotypes. *Journal of Personality and Social Psychology*, 1973, 4, 75-88.
- Hill, C. T., Rubin, Z., & Peplau, A. Breakups before marriage: The end of 103 affairs. *Journal of Social Issues*, 1976, 32, 147-168.
- Karlsson, G. *Adaptability and communication in marriage: A Swedish predictive study of marital satisfaction*. Uppsala, Sweden: Almqvist & Wiksells, 1951.
- Kelly, E. L. Concerning the validity of Terman's weights for predicting marital happiness. *Psychological Bulletin*, 1939, 36, 202-203.
- Kernodle, W. Some implications of the homogamy-complementary needs theories of mate selection for sociological research. *Social Forces*, 1969, 38, 145-152.
- Landis, J. T. The pattern of divorce in three generations. *Social Forces*, 1956, 34, 213-216.
- Landis, J. T. Social correlates of divorce or non-divorce among the unhappily married. *Marriage and Family Living*, 1963, 25, 178-180.
- Laws, J. A feminist review of marital adjustment literature: The rape of the Locke. *Journal of Marriage and the Family*, 1971, 33, 483-516.
- Locke, H. J. *Predicting adjustment in marriage: A comparison of a divorced and a happily married group*. New York: Holt, 1951.
- Locke, H. J., & Karlsson, G. Marital adjustment and prediction in Sweden. *American Sociological Review*, 1952, 17, 10-17.
- Locke, H. J., & Wallace, K. M. Short marital adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living*, 1959, 21, 251-255.
- Luckey, E. B. Number of years married as related to personality perceptions and marital satisfaction. *Journal of Marriage and the Family*, 1966, 28, 44-48.

- Meyer, J. P., & Pepper, S. Need compatibility and marital adjustment in young married couples. *Journal of Personality and Social Psychology*, 1977, 35, 331-342.
- Murstein, B. J. The complementary need hypothesis in newlyweds and middle-aged couples. *Journal of Abnormal and Social Psychology*, 1961, 63, 194-197.
- Murstein, B. J. Empirical tests of role, complementary needs and homogamy theories of marital choice. *Journal of Marriage and the Family*, 1967, 29, 689-696.
- Murstein, B. J., & Beck, G. D. Person perception, marriage adjustment and social desirability. *Journal of Consulting and Clinical Psychology*, 1972, 39, 396-403.
- Murstein, B. J., & Glaudin, V. The relationship of marital adjustment to personality: A factor analysis of the Interpersonal Checklist. *Journal of Marriage and the Family*, 1966, 28, 37-43.
- Olkin, I., & Pratt, J. W. Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 1958, 29, 201-211.
- Pickford, J. H., Signori, E. J., & Rempel, H. The intensity of personality traits in relation to marital happiness. *Journal of Marriage and the Family*, 1966, 28, 458-459. (a)
- Pickford, J. H., Signori, E. J., & Rempel, H. Similar or related personality traits as a factor in marital happiness. *Journal of Marriage and the Family*, 1966, 28, 190-192. (b)
- Pope, H., & Mueller, C. W. The intergenerational transmission of marital instability: Comparisons by race and sex. *Journal of Social Issues*, 1976, 32, 49-66.
- Reevy, W. R. Vestured genital apposition and coitus. In H. G. Beigel (Ed.), *Advances in sex research*. Harper & Row, 1963.
- Renne, K. S. Health and marital experience in an urban population. *Journal of Marriage and the Family*, 1971, 33, 338.
- Rosow, I. Issues in the concept of need complementarity. *Sociometry*, 1957, 20, 216-233.
- Singh, S. B., Nigam, A., & Saxena, N. K. 16 P.F.: Study in the cases of marital disharmony. *Indian Journal of Clinical Psychology*, 1976, 3, 47-52.
- Terman, L. M., & Oden, M. H. *The gifted child grows up: Twenty-five years' follow-up of a superior group*. Stanford, Calif.: Stanford University Press, 1947.
- Tharp, R. J. Psychological patterning in marriage. *Psychological Bulletin*, 1963, 60, 97-117.
- Udry, J. R. Personality match and interpersonal perception as predictors of marriage. *Journal of Marriage and the Family*, 1967, 29, 722-725.
- U.S. Bureau of the Census, 1970 Census of Population. *Age at first marriage* (Final Rep. PC(2)-4D). Washington, D.C.: U.S. Government Printing Office, 1973.
- Weigel, R. G., Weigel, V. M., & Richardson, F. C. Congruence of spouses' personal constructs and reported marital success: Pitfalls in instrumentation. *Psychological Reports*, 1973, 33, 212-214.
- Wiggins, J. S. *Personality and prediction: Principles of personality assessment*. Reading, Mass.: Addison-Wesley, 1973.
- Winch, R. F. Another look at the theory of complementary needs in mate selection. *Journal of Marriage and the Family*, 1967, 29, 756-767.
- Woodward, J. A., & Bentler, P. M. A statistical lower-bound to population reliability. *Psychological Bulletin*, in press.

Received August 9, 1977 ■

Sum of MMPI Scales *F*, 4, and 9 as a Measure of Aggression

L. Rowell Huesmann

University of Illinois at Chicago Circle

Monroe M. Lefkowitz

New York State Department of Mental Hygiene
Albany, New York

Leonard D. Eron

University of Illinois at Chicago Circle

The relationship between scores on the Minnesota Multiphasic Personality Inventory and both concurrent and prior aggression was examined for a sample of 426 19-year-olds from the general population. Aggression was measured through peer nominations obtained concurrently and 10 years earlier. Correlation and regression analysis indicated that the sum of *T* scores for Scales *F*, 4, and 9 was a valid measure of aggression. The composite was also shown to have a higher reliability than its component scales. Using an additional 283 subjects from delinquent populations, it was demonstrated that the composite was an excellent discriminator between delinquent and general populations of males and females even when intelligence and social status were controlled.

Elevation of the *T* scores on both Scales 4 and 9 of the Minnesota Multiphasic Personality Inventory (MMPI) has been thought to form the profile characteristic of the male juvenile delinquent (Dahlstrom & Welsh, 1960; Dahlstrom, Welsh, & Dahlstrom, 1972; Hathaway & Monachesi, 1953). Scale 4 (Psychopathic Deviate) by itself has been used to measure levels of social deviance or antisocial behavior (Elion & Megargee, 1975; Hathaway & Monachesi, 1953; Megargee & Mendelsohn, 1962). Dahlstrom et al. (1972) noted that marked elevations on Scale 4 can be observed in prison groups. These authors also noted that Scale 9 (Hypomania) appears to energize the pattern related to Scale 4. Scale 9 was viewed by Hathaway and Monachesi (1953) as an exciter that in combination with

Scale 4 produces rebellious and excitable behavior in high-delinquent children. A later analysis of their data (Monachesi & Hathaway, 1969) showed that the highest rates of delinquency for both boys and girls were associated with deviance in MMPI Scales 4, 8, and 9.

In a study of juvenile delinquents who succeeded or failed in adjustment to institutionalization, Lefkowitz (1966) found that the mean score on Scale 9 was significantly higher for the failures. Similarly, in a follow-up study of paroled prisoners whose postrelease behavior was classified as acceptable or unacceptable, the MMPI 49 code type was heavily represented among the unacceptable group (Jacobson & Wirt, 1969). Butcher (1965) found that highly aggressive boys (based on peer nominations) had significant elevations on Scales 4 and 9. These boys responded in a rebellious and excitable manner in interpersonal situations, whereas the low-aggressive boys tended to internalize their conflicts, which were then manifested in hypochondriacal symptoms and withdrawal.

There has been a proliferation of attempts to validate not only certain of the clinical scales as aggression measures (see, e.g., Megargee & Mendelsohn, 1962; Shipman, 1965) but also a number of special scales developed from the MMPI. At this writing,

This study was supported by Grant M1726 from the National Institute of Mental Health, by Contract HSM 42-70-6 from the Surgeon General's Scientific Advisory Committee on Television and Social Behavior, and by the New York State Department of Mental Hygiene. The data analysis was performed with the assistance of the computation center of the University of Illinois at Chicago Circle. For Part 1 of this study, Leopold Walder was a co-investigator.

Requests for reprints should be sent to L. Rowell Huesmann, Department of Psychology, University of Illinois at Chicago Circle, Box 4348, Chicago, Illinois 60680.

as many as 25 scales of the MMPI have been proposed as measures of hostility or aggression (Deiker, 1974). Thus, the present study is an attempt to simplify this area of measurement by providing a reliable and valid single score of aggressive behavior obtained by summing the *T* scores of a few common clinical scales. The virtue of a single score over a profile of scale scores is that the former is readily usable as a continuous variable, whereas the latter is a dichotomous variable and therefore in the main limited to frequency analysis.

A two-step procedure was followed in this study. First, using a sample of subjects from a general population, we attempted to identify a valid and reliable composite of *T* scores for measuring aggression. Then using a sample of subjects from a delinquent population, we attempted to validate the composite as a discriminator of delinquent adolescents.

Study 1

Method

Subjects. The subjects were part of a larger longitudinal research project on aggressive behavior reported elsewhere (Eron, Huesmann, Lefkowitz, & Walder, 1972; Lefkowitz, Eron, Walder, & Huesmann, 1977). In the first wave, data were gathered during 1959-1960 from the entire 3rd-grade population of 875 boys and girls residing in a semirural county in New York State. Ten years later in the second wave, termed the "13th grade," data were collected from 426 of these subjects (211 boys and 215 girls who could be located at that time; Lefkowitz et al., 1977). In the 3rd and 13th grades, the modal ages of this sample of 426 were 8 and 19 years, respectively. The mean IQ of this sample in the 3rd grade was 107.10 ± 13.66 . Based on fathers' occupation, the sample can be described as predominantly middle class.

Procedure. In the 3rd and 13th grades, aggression scores were obtained by a peer nomination questionnaire comprised of items describing aggression behavior. For example, each subject was asked to name anyone in the classroom "Who starts a fight over nothing?" The peer nomination technique also included items intended to measure aggression avoidance; for example, "Who will never fight even when picked on?" The reliability ($r > .85$) and validity of this aggression measure and its the procedures for administration have been discussed elsewhere (Butcher, 1965; Eron, Walder, & Lefkowitz, 1971; Walder, Abelson, Eron, Banta, & Laulicht, 1961). In the 3rd grade the peer nomination technique, including 10 aggression items, was administered to classroom groups. In the 13th grade the technique, using 9 of the 10 aggression items from the 3rd grade, was administered individually to subjects as part of a larger 2-hour interview procedure. At the time of the 3rd grade interview, IQ (Sullivan, Clark, & Tiegs, 1957) and father's occupational status (U.S. Bureau of Census, 1960) were recorded for most subjects. During the 13th grade interview, subjects also completed self-report inventories on aggressive behavior and were administered the MMPI. The *K*-corrected *T* scores on the MMPI were used as the potential predictors of aggression in this study. The self-report measure of aggression, termed *total aggressive environment*, was comprised of 43 items divided into five subscales: (a) respondent as a victim of aggression, (b) respondent as a witness of aggression, (c) respondent's aggressive habits, (d) respondent's antisocial behavior, and (e) respondent's aggressive feelings. The exact composition of this scale and the procedures used to administer all of the 13th grade measures have been reported elsewhere (Lefkowitz et al., 1977).

Results

The intercorrelations between the various measures of aggression are shown in Table 1 for all 426 subjects. Since 13th grade peer-rated aggression (Peer Agg 13) correlated well with all the other measures and was measured at the same time that the MMPI was admin-

Table 1
Correlations Between the Criterion Measures of Aggression for All 426 Subjects

Measure	1	2	3	4	5
1. Peer Agg 13	—				
2. Peer Avoid Agg 13	-.364	—			
3. Self Rep Agg 13	.520	-.357	—		
4. Peer Agg 3	.420	-.284	.200	—	
5. Peer Avoid Agg 3	-.236	.298	-.233	-.381	—
6. Sex ^a	.346	-.080	.390	.199	-.139

Note. $r_{.001} = .15$. Peer = peer nominated; Agg = aggression; Avoid = avoidance of; Rep = reported; 3 refers to Grade 3; 13 refers to Grade 13.

^a Female = 0; male = 1.

Table 2
*Correlations and Significant Multiple Regression Coefficients (MRs) for Predicting
 Concurrent Aggression from Minnesota Multiphasic Personality Inventory (MMPI) Scales*

MMPI scale	Concurrent peer-rated aggression			
	Males ^a		Females ^b	
	<i>r</i>	Raw regression coefficient	<i>r</i>	Raw regression coefficient
<i>L</i>	.064		.087	
<i>F</i>	.322***	1.675*	.260***	1.302***
<i>K</i>	-.134		-.056	
1	.052		.049	
2	.077		.005	
3	.001		.090	
4	.339***	2.469***	.327***	1.073***
5	-.129	-1.799**	.182**	.561*
6	.180**		.121	
7	.131		-.017	
8	.232***		.111	
9	.332***	1.351*	.146*	
10	-.132		-.150*	-.674*

^a For males, $n = 211$, $R = .471^{***}$.

^b For females, $n = 215$, $R = .478^{***}$.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

istered, it was decided to use it as the primary criterion measure. Since sex was significantly related to aggression, all further analysis were done separately for males and females.

Table 2 shows the correlations between Peer Agg 13 and each of the MMPI scales as well as the results of a multiple regression analysis predicting Peer Agg 13 from the MMPI scales. For males, as expected, Scales 4 and 9 were highly significant correlates of aggression, but Scale *F* was also highly correlated. Scale 8 correlated less highly though significantly. For females, the two best univariate predictors were Scale 4 and the *F* scale. Scales 5, 9, and 10 also were significantly correlated with aggressiveness but not as highly. Since the *T* scores for Scales 4, 7, 8, and 9 were *K* corrected and since *K* scores were slightly negatively correlated with aggressiveness, the observed correlations between aggressiveness and 4, 7, 8, and 9 may have been slightly reduced by the *K* correction. However, the reduction could not be significant considering the small size of the correlation between *K* and Peer Agg 13.

The multiple regression analysis shows the

best weighted composite of the *T* scores for measuring aggression.¹ As expected, the regression coefficients reveal that Scale 4 was the most important predictor of aggression across sexes. Surprisingly, however, Scale *F* was as important a predictor for females and equal with Scales 5 and 9 in predicting aggression in males. Scale 5 relates negatively to aggression in males and positively in females, indicating the relation between aggression and sex role behaviors. Scale 9 made an important contribution for males but not for females, for whom Scale 10 appeared as an inverse predictor.

Discussion

From these results it appears that a composite to be used as a general measure of

¹ The raw regression coefficients are presented because the objective is to develop a predictive composite. Even though the intercorrelations between the MMPI scales are mostly significant, they only range from $-.32$ to $.52$ for the significant scales in the equation, so multicollinearity should not be a serious problem.

Table 3
Correlations Between the Composite Minnesota Multiphasic Personality Inventory Scale, Its Components, and the Measures of Aggression

Aggression measure	Males (<i>n</i> = 211)						Females (<i>n</i> = 215)			
	Scale									
	F + 4 + 9	F + 4	F	4	9	F + 4 + 9	F + 4	F	4	9
Concurrent										
Peer Agg 13	.414***	.377***	.322***	.339***	.332***	.313***	.343***	.257***	.327***	.146*
Peer Avoid Agg 13	-.314***	-.274***	-.252***	-.227***	-.271***	-.257***	-.246***	-.179**	-.241***	-.177**
Self-Rep Agg 13	.495***	.433***	.359***	.401***	.426***	.408***	.354***	.200**	.393***	.342***
10 years prior										
Peer Agg 3	.219***	.218***	.170*	.212***	.144*	.137*	.107	.101	.086	.134*
Peer Avoid Agg 3	-.311***	-.303***	.248***	.284***	-.216***	-.079	-.085	-.100	-.052	-.038

Note. Peer = peer nominated, Agg = aggression; Avoid = avoidance of; Rep = reported; 3 refers to Grade 3; 13 refers to Grade 13.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

aggression across sexes without contamination by sexual stereotyping should be composed of Scales *F* and 4 and either Scale 9 or scale 10. Of course, a weighted linear composite involving every scale would have the highest correlation with aggression. However, a measure consisting of the sum of only two or three scales has the advantage of simplicity and ease of administration. Furthermore, as Wainer (1976) and others have argued, the simple sum of significant predictor variables is often almost as good a predictor in the population as the entire multiple regression equation.

Scale 5, although a significant predictor, is not a desirable variable to include in a general measure of aggression, because it represents that component of aggression associated with sexual stereotyping. Since it was included in the regression equation, however, the effects of sexual stereotyping have been partially controlled in determining the other significant predictors.

On a priori grounds, Scale 9 would seem to be a more appropriate element to include in a measure of aggression than Scale 10. It is intended to measure a propensity for manic behavior, so it certainly has face validity as a component of aggression. Furthermore, as reported in the Introduction, Scale 9 has been found to be significantly higher in several high-aggressive, albeit male, populations. On the other hand, Scale 10 has to our knowledge never been proposed as a discriminator of delinquents. In addition, the utility of Scale 10 in this study could possibly be an artifact of the peer nomination procedure. A female scoring high on Scale 10 is scoring high on social introversion. Therefore, she might be less likely to be nominated by her peers on any scale resulting in negative correlations. The available data are consistent with this interpretation. Even though Scale 10 correlated negatively with peer-rated aggression for females ($r = -.15$, $p < .03$), it did not correlate significantly with self-rated aggression ($r = .04$), though one would usually expect a higher correlation between two self-ratings. Also, peer-rated popularity, which correlated negatively with peer-rated aggression ($r = -.28$, $p < .001$), had exactly the same correlation with Scale 10 ($r = -.15$) as did peer-rated aggression. Thus, it seems that

Scale 10 is measuring a component of peer rating more than aggression. Scale 9, on the other hand, correlated more significantly with self-rated aggression for girls ($r = .34$, $p < .001$) than with peer-rated aggression ($r = .15$, $p < .03$). Therefore, a composite measure of aggression was constructed consisting of the sum of Scales *F*, 4, and 9.

Table 3 displays the correlations between this composite measure of aggression and each of the other variables measuring aggression. It also contains the correlations between the aggression measures and the individual scales. One can see that for males the sum of Scales *F*, 4, and 9 is a valid measure of concurrent aggression and is significantly related to the subject's aggression to years earlier. The composite correlated more strongly with every measure than any of its components or *F* + 4. For females, the composite also correlated significantly with every measure of concurrent aggression, though less strongly than for males. Scale 4 alone and *F* + 4 are just as strongly or slightly more strongly related to peer ratings than the full composite, but the full composite did better on the other measures. The composite also significantly correlated with aggression 10 years earlier.

Reliability

The reliability of any linear combination of variables can be computed from its variance and the variances and reliabilities of its components. For the current sample, however, MMPI item scores were not available to us, so reliability data from comparable samples had to be used. The largest study of reliability in a college-age population of normals appears to have been conducted by Mauger in 1972 (cited in Dahlstrom et al., 1972). Test-retest correlations were computed for 490 subjects over an 8-month lag. The average correlations of males and females were .56 on Scale *F*, .57 on Scale 4, and .62 on Scale 9. These relatively long-term stability coefficients can be viewed as lower bounds on internal consistency reliabilities (i.e., coefficient alpha). Dahlstrom et al. (1972) did not report any studies involving a substantial number of normal college-age subjects for which coefficient alpha was calculated, so these stability coefficients

Table 4

Means for F + 4 + 9 as a Function of Sex and Delinquency

Population	Males	Females	Total
General	183.3	174.6	178.9
<i>n</i>	210	215	
Delinquent	217.4	237.7	227.2
<i>n</i>	147	136	
Total	197.3	199.0	198.2

will have to serve as our conservative estimates of reliabilities. We can now show that even assuming such conservative estimates, the reliability of *F* + 4 + 9 is acceptable.

From Nunnally (1967, p. 229), we can write:

Let

$$y = \text{MMPI}_F + \text{MMPI}_4 + \text{MMPI}_9$$

$$r_{yy} =$$

$$1 - \frac{(\sigma_F^2 + \sigma_4^2 + \sigma_9^2) - r_{FF}\sigma_F^2 - r_{44}\sigma_4^2 - r_{99}\sigma_9^2}{\sigma_y^2}$$

Using our sample of 427 subjects to estimate the standard deviations gives

$$\sigma_y = 26.52 \quad \sigma_F = 10.82 \quad \sigma_4 = 11.15 \quad \sigma_9 = 11.50,$$

and using Mauger's stability coefficients as conservative estimates of reliability gives

$$r_{FF} = .56, \quad r_{44} = .57, \quad \text{and} \quad r_{99} = .62.$$

Therefore,

$$r_{yy} = 1 - \frac{155.23}{703.31} = .78.$$

Thus the reliability of *F* + 4 + 9 is substantially higher than the reliabilities of its components and sufficient for its use as a measure of aggression. Furthermore, using less conservative estimates of the reliabilities of the scales based on 1-week stabilities (Dahlstrom et al., 1972), the reliability of *F* + 4 + 9 is almost .87.

Study 2

To further validate *F* + 4 + 9 as a measure of aggression, a second analysis was undertaken to compare a sample from a known population of delinquents with the previously studied sample from a normal population.

Table 5

Analysis of Covariance for $F + 4 + 9$ as a Function of Sex and Delinquency

Source	df	SS	MS	F
Covariates				
IQ	1	157,298.44	157,298.44	218.27*
Father's occupation	1	13,772.81	13,772.81	19.11*
Effects				
Sex (A)	1	10,496.69	10,496.69	14.57*
Population (B) ^a	1	153,635.63	153,635.63	213.19*
A \times B	1	17,628.37	17,628.37	24.46*
Residual	621	447,522.25	720.65	
Total	626	800,354.19		

Note. The total sample size for the analysis of covariance was only 627 because IQ or father's occupation was missing for 81 subjects.

^a Delinquent versus general.

* $p < .001$.

Method

Subjects. The delinquency sample consisted of 136 females institutionalized at a facility of the New York State Division for Youth and 147 males from a privately operated institution for delinquent boys in New York.² All subjects had been sent to the institutions by family and children's courts. The subjects ranged in age from 12.9 to 17.1, with a median of 14.9. The mean IQ for girls was only 77.3 but was 100.3 for boys. The entire population of the institutions minus a few subjects for whom usable data could not be obtained constituted the sample. The offenses committed by the delinquents ran the gamut from incorrigibility, shoplifting, petty larceny, and prostitution to car theft, breaking and entering, and assault.

Procedure. Form R of the MMPI was group administered to 3-5 subjects at a time. The 399 items were read aloud twice on a tape while the subjects, closely monitored, read along silently in their booklets. The IQs and fathers' occupations were obtained from the subjects' case histories.

Results

The means and analysis of covariance shown in Tables 4 and 5 indicate that $F + 4 + 9$ is an excellent discriminator between populations known to vary in their levels of aggression. The analysis of covariance was performed in a hierarchical manner, so the F value for delinquent versus nondelinquent populations, $F(1, 621) = 244.1$, $p < .001$, represents the discriminative strength of $F + 4 + 9$ after the effects of IQ, father's occupation, and sex have been partialled out. Since the standard deviation of the composite was about 25, the table of means indicates that delinquent boys scored about $1\frac{1}{2}$ standard deviations higher

than normals, and delinquent girls scored about $2\frac{1}{2}$ standard deviations higher. This greater difference for females is reflected in the significant Sex \times Population interaction $F(1, 621) = 25.83$, $p < .001$. Although males scored higher than females on $F + 4 + 9$ in the normal population, $t(423) = 3.42$, $p < .001$, females scored higher in the delinquent population, $t(281) = 5.98$, $p < .001$.

Discussion

These results provide construct validity for the sum of MMPI Scales F , 4, and 9 as a measure of aggression. The sum of these scales was an excellent discriminator between delinquents high in aggression and youth from a normal population even when IQ and social class were controlled. Furthermore, as one would predict for a measure of aggression, nondelinquent males scored significantly higher on $F + 4 + 9$ than nondelinquent females. Within the delinquent group, however, females, contrary to the hypothesis, scored significantly higher than males. Speculatively, this unexpected result of higher scores for delinquent females than delinquent males on this MMPI measure may be due to an artifact of selection. Traditionally, females have encountered far fewer difficulties with law-enforcing agencies

² The authors wish to express their thanks for the cooperation of the staffs of the Hudson School for Girls and the Berkshire Farm Institute for Training and Research.

and/or have been treated more leniently by these agencies than males, particularly juveniles (Monachesi & Hathaway, 1969). Thus, in order for a female to be considered delinquent and institutionalized, the level of antisocial behavior may have to be excessive as compared to males. Consequently, the present group of institutionalized delinquent females may be unrepresentative of delinquent females generally, in that their level of aggressiveness may be inordinately high. The unusually low mean IQ for the female delinquents supports this interpretation. If this post hoc hypothesis is valid, this MMPI measure should be able to discriminate among degrees of delinquency in a population of delinquent females.

Although the normals in the present study were approximately 5 years older than the delinquents, it is unlikely that this age difference affected the results. Hathaway and Monachesi (1963) provided MMPI norms for 4,944 boys and 5,207 girls in the ninth grade whose average ages, respectively, were 15.1 and 14.9 years. The sum of the *T* scores on Scales *F*, 4, and 9 were 176.3 for boys and 170.4 for girls. These values are markedly below those of delinquents of the same age (as shown in Table 4) but approximately the same as those of the normal subjects in the present study.

Summary

The results of the present study indicate that the sum of the *T* scores on MMPI Scales *F*, 4, and 9 serves as a reliable and valid unitary measure of aggression. Highly significant correlations obtained between *F* + 4 + 9 and both peer-rating and self-rating of aggression. Added strength is lent to the validity by the fact that scores on the measure related back 10 years in time to nominations on aggressive behavior that the subjects received from their peers in a classroom setting. Moreover, as evidenced by the multiple regression analysis, the composite is valid for both males and females and is not simply a sex-typing measure. In general, the sum of these scales appeared more valid than any of its components. Other support for the validity of this MMPI measure was derived from its ability

to distinguish between delinquent and non-delinquent populations known to differ in aggressiveness. Finally, *F* + 4 + 9 was shown to have a reliability greater than that of any of its components.

Although the intent was to validate this measure on a noninstitutionalized population of normal subjects, the fact that the measure discriminates so well between institutionalized and noninstitutionalized populations suggests that it may also be useful in criminal justice settings in which the MMPI is so widely used. The ease with which this measure is obtained from the MMPI makes it a potentially useful tool for screening and perhaps program placement in such settings. Also, the measure meets the demand for a valid paper-and-pencil measure of aggressive behavior for research with normal subjects.

References

- Butcher, J. Manifest aggression: MMPI correlates in normal boys. *Journal of Consulting Psychology*, 1965, 29, 446-454.
- Dahlstrom, W. G., & Welsh, G. S. *An MMPI handbook*. Minneapolis: University of Minnesota Press, 1960.
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook*. Minneapolis: University of Minnesota Press, 1972.
- Deiker, T. E. A cross-validation of MMPI scales of aggression on male criminal criterion groups. *Journal of Consulting and Clinical Psychology*, 1974, 42, 196-202.
- Elion, V. H., & Megargee, E. I. Validity of the MMPI *Pd* scale among black males. *Journal of Consulting and Clinical Psychology*, 1975, 43, 166-172.
- Eron, L. D., Huesmann, L. R., Lefkowitz, M. M., & Walder, L. O. Does television violence cause aggression? *American Psychologist*, 1972, 27, 253-263.
- Eron, L. D., Walder, L. O., & Lefkowitz, M. M. *Learning of aggression in children*. Boston: Little, Brown, 1971.
- Hathaway, S. R., & Monachesi, E. D. *Analyzing and predicting juvenile delinquency with the MMPI*. Minneapolis: University of Minnesota Press, 1953.
- Hathaway, S. R., & Monachesi, E. D. *Adolescent personality and behavior: MMPI patterns of normal, delinquent, dropout, and other outcomes*. Minneapolis: University of Minnesota Press, 1963.
- Jacobson, J. L., & Wirt, R. D. MMPI profiles associated with outcomes of group psychotherapy with prisoners. In J. N. Butcher (Ed.), *MMPI: research developments and clinical applications*. New York: McGraw-Hill, 1969.
- Lefkowitz, M. M. MMPI scores of juvenile delinquents adjusting to institutionalization. *Psychological Reports*, 1966, 19, 911-914.

- Lefkowitz, M. M., Eron, L. D., Walder, L. O., & Huesmann, L. R. *Growing up to be violent: A longitudinal study of the development of aggression*. New York: Pergamon Press, 1977.
- Megargee, E. I., & Mendelsohn, G. A. A cross-validation of twelve MMPI indices of hostility and control. *Journal of Abnormal and Social Psychology*, 1962, 65, 431-438.
- Monachesi, E. D., & Hathaway, S. R. The personality of delinquents. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications*. New York: McGraw-Hill, 1969.
- Nunnally, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.
- Shipman, W. G. The validity of MMPI hostility scales. *Journal of Clinical Psychology*, 1965, 21, 186-190.
- Sullivan, E. T., Clark, W. W., & Tiegs, E. W. *California short-form test of mental maturity*. Los Angeles: California Test Bureau, 1957.
- United States Bureau of Census. *Classified index of occupations and industries*. Washington, D.C.: U. S. Government Printing Office, 1960.
- Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 1976, 83, 213-217.
- Walder, L. O., Abelson R., Eron, L. D., Banta, T. J., & Laulicht, J. H. Development of a peer-rating measure of aggression. *Psychological Reports*, 1961, 9, 497-556.

Received August 19, 1977

A Longitudinal Study of Coping Styles in Self-Defining and Socially Defined Women

Abigail J. Stewart
Boston University

In a longitudinal study of 51 female college graduates, self-definition as measured by freshman-year Thematic Apperception Tests predicted several aspects of problem-solving and coping behavior 14 years later. Those women who viewed themselves, their world, and their own personal problems in ways that facilitated effective coping scored higher in self-definition. The women who took instrumental actions to solve their problems, rather than taking noninstrumental actions or remaining passive, also scored higher in self-definition. Theoretical and practical implications of these findings are discussed.

Self-definition and social definition were proposed by Stewart and Winter (1974) as two contrasting patterns of organization of experience that are important in understanding intrasex differences in women.¹ Stewart and Winter began by identifying two different styles of storytelling that characterized Thematic Apperception Test (TAT) stories of career-oriented and non-career-oriented female college students. The stories written by the career-oriented women were marked by a clear, causally organized plot and instrumentally active and effective characters. The stories written by the non-career-oriented women were characterized by inactivity or ineffective activity on the part of characters, and by an absence of intelligible causal ordering of the plot.²

Stewart and Winter argued that these two styles of storytelling reflected an underlying pervasive personality characteristic. They presented evidence that the women who told stories of rationally caused events created by

effective actors were (a) relatively indifferent to sex role norms (and perhaps to norms associated with other roles); (b) capable of emotional "distance" and objectivity in the context of a small number of intense, personal relationships; (c) active and interested in broad social movements and issues; and (d) capable of vigorous instrumental activity.

The women who told stories in which events were chaotic and in which characters were helpless (a) behaved in accord with sex role norms; (b) showed rather superficial emotional attachments to a large number of people, with little capacity for analytic "distance"; (c) were deeply preoccupied with their own emotional and social lives; and (d) showed a greater inclination toward expressive rather than instrumental activity. The initial research reported by Stewart and Winter (1974), then, indicated that these two patterns of thought (self- and social definition) were coherently reflected in two similar patterns of ordinary behavior in female college students. In later research, Stewart (1975) showed that self-definition predicted career and work activities as well as certain marital and family patterns in adult women. Winter, Stewart, and Mc-

The author wishes to thank David C. McClelland and David G. Winter for their advice and help with this research, which was partially supported by a fellowship from the Peter B. Livingston Foundation and a grant from the Fund for the Improvement of Post-Secondary Education.

Requests for reprints should be sent to Abigail Stewart, Department of Psychology, Boston University, 64 Cummington Street, Boston, Massachusetts 02215.

¹ Later research (see Winter, Stewart, & McClelland, 1977) has indicated that it may also be relevant in studying men.

² Validation research and the psychometric properties of the measure are discussed in detail in Stewart and Winter (1974).

Clelland (1977) also showed that self-definition was associated with nontraditional marriage patterns in a longitudinal study of adult men.

No research, however, has explored the possibility that the patterns associated with self- and social definitions might extend to the domain of adult behavior in response to life stress. Nevertheless, the presumably general and stable tendencies to organize experience identified in Stewart and Winter's (1974) original study should have important consequences for an individual's approach to solving a personal problem. A great deal of theory and laboratory evidence suggests that individuals' attributions about the causes of events (see de Charms, 1968; Kelley, 1972) and their feelings of personal control over events (see Rotter, 1966; Seligman, 1975) have important consequences both for their acts and their emotional states. Self- and social definition are alternative patterns of organization of experience that do not involve *conscious beliefs* (as "locus of control" does) but rather *implicit beliefs*. Moreover, the two patterns simultaneously reveal implicit beliefs about the causal organization of external reality and one's own capacity for causal agency (see, e.g., de Charms, 1968) rather than being restricted to one or the other. It is assumed here that stable patterns of thought (implicit belief systems) about one's capacity to influence events, and about the predictability and meaningfulness of events in general, will affect individuals' styles of responding to personal life events.

Thus, if a woman sees the world as a systematic environment in which events have identifiable causes, it seems reasonable that she would seek out the causes of her problems. If, however, a woman sees the world as an essentially random environment in which events lack intelligible causes, she will waste no time in fruitless searches for nonexistent explanations for her problems. Also, a woman who sees the individual as a locus of effective voluntary action will take action to solve her problems, whereas a woman who sees individual action as futile and pointless may simply wait for this, too, to "pass."

On the other hand, the link between world view and coping style may not be so close. A socially defined woman could "typically"

see the world as chaotic and herself as helpless, but when faced with real problems she might seek out the causes of the problems and rational action to solve them. Alternatively, a self-defining woman who "normally" sees the world as orderly and herself as effective might be overwhelmed and paralyzed by a personal difficulty. One test of an individual's commitment to a world view, then, may be the extent to which that world view is retained in the face of calamity.

The purpose of this study was to test the general hypothesis that self-defining and socially defined women would indeed respond to personal life stresses differently and in ways that are consistent with their respective world views. An adequate test of this hypothesis offers the possibility of extending the construct validity of self-definition and at the same time determining whether alternative coping styles are grounded in characteristic constructions of social reality.

Hypotheses

The first hypothesis investigated in this study is that self-defining women will be more likely than socially defined women to interpret their own personal problems as existing outside of themselves. That is, they will tend to see a problem (regardless of its content) as located in the environment, rather than within their own personality. However, the second hypothesis is that self-defining women will be more likely than socially defined women to interpret the *solution* to their problems as located within themselves. That is, they will see the *problem* as external to themselves, but the *solution* will be seen as internal; in this way, it is presumed, they may reasonably plan to initiate action to solve the problem. If the problem seems internal, or if the solution of the problem seems external, it may make little sense to take action. Only if the problem appears to be external, but the solution to the problem appears to be within one's control, is it reasonable for an individual to take action to solve it.

In addition, it is hypothesized that self-defining women will be better able than socially defined women to clearly articulate the *nature* of their problems and their solutions,

and to identify the causes of these problems. That is, the stable tendency to see the world as rational and causally organized will be reflected in a tendency to see a particular problem as the result of a number of factors that are intelligible and identifiable.

It is further hypothesized that self-defining women will be more likely to present their problems in a broad social context. They will be more likely to experience their problems as involving a wider network of social relationships (e.g., colleagues in work and voluntary settings, the wider society), whereas socially defined women will be inclined to define their problems as located within their personal domestic context (i.e., the home, marriage, or family).

Finally, it is hypothesized that self-defining women will be more likely than socially defined women to make instrumental responses to their problems, and that socially defined women will be more likely to make either noninstrumental responses or no active responses at all to their problems. Thus, overall, it is hypothesized that an individual's particular interpretation of and response to a particular life problem may be grounded in a dispositional world view.

In order to examine these substantive hypotheses it is, however, also necessary to determine whether self-definition is associated with the *type of problem* that a woman faces. That is, if we wish to conclude that interpretive differences between the two groups are meaningful, and that the differences are a function of the women's viewpoints, we must show that the differences are *not* a function of the actual problems faced by the two groups.

Method

Subjects and Procedure

The subjects in this study were 60 women, randomly selected from a larger longitudinal study of 122 women educated at an elite New England women's college. A six-picture TAT had been administered to the women under neutral conditions in the fall of their freshman year of college, as part of a larger study of college students. Ten years after their graduation from college (when they were 31 years old), they were mailed a questionnaire, inquiring about their activities in the years since college. In addition, all of the women were asked whether they would be willing to be interviewed. Because only 30% of the sample lived within a distance

permitting a personal interview, a random sample of 60 of the 122 women was selected to be interviewed over the phone. Of the 60 women, 57 were successfully contacted by telephone and interviewed for 20–30 minutes. The interviews were taped, with subjects' knowledge and permission, and were later transcribed for coding, with any identifying indications deleted. Complete, codable transcriptions were made for 51 of the 57 interviewees. The remaining 6 interviews were at least partially inaudible or unintelligible due to various mechanical failures.

Female interviewers³ conducted the interviewing according to the interview schedule outlined below, mainly attempting to ensure that in discussing the material, subjects covered all topics in the schedule. Subjects were surprisingly unruffled by the prospect of being interviewed over the telephone, and the resulting data appear quite open and complete. It is possible that the relative anonymity and impersonality of a telephone conversation encourages free and open conversation, since there is little opportunity for inadvertent expression of disapproval or disagreement. The hypothetical disadvantages of telephone interviews (lack of personal contact, lower rapport, etc.) did not seem to be significant factors in this study.

The Interview Schedule

Interviewers read the following paragraph to the subjects:

We are interested in studying stress in women's lives. We would like you to describe the period in your life that you think of as the most unhappy or upsetting time that you've lived through. Any period is all right, but choose a time when—for an extended period, not just briefly—things really seemed to go badly. The time needn't be catastrophic—just a time when things were not going well.

If subjects responded with a question about how long was "an extended period," interviewers were instructed to indicate that "the period should be about 6 months or more." If subjects indicated that nothing really terrible had happened, interviewers were instructed to respond, "just pick a time when you were less happy and more frustrated or distressed than usual." If subjects indicated that more than one period would qualify, interviewers asked them to select the period they felt was "most stressful for the longest period." In addition, interviewers used the following probes if the subjects' responses did not answer the questions involved: "What led up to this situation?" "What were the causes of this situation?" "Why did you find these particular things so upsetting?" "What did you do?" "Did you take any steps to deal with your situation?" "How did this all end?"

At the end of the interview, the interviewer concluded: We feel that women's lives are especially stressful, but we hope to understand these stresses better, so

³ The author is grateful to Betsy Harrington and Kathleen Finn for their assistance as interviewers.

we can help do something about them. Would you like a summary of our findings?

The interviewer then filled out a card with the subject's name and address, and feedback was later sent to the subject.

Coding the Data

I scored the original TATs, administered in 1960, blindly for self-definition before the interviews were conducted. I have demonstrated interrater reliability with other scorers on this variable ($\rho = .95$). In addition, I have demonstrated score-rescore and intercoder reliabilities (ρ s) above .94 on a subsample of the data scored for this study.

Briefly, the scoring categories in the self-definition scoring system⁴ can be summarized as follows:

Self-definition categories (scored +1). Causality: Use of explicit causal language. Reason-action sequence: The plot is organized in such a way that the final element of the story is an action or plan for action which follows logically from the preceding elements.

Social definition categories (scored -1). No causality: Events occur that appear to have no known causes or that are bizarre and unintelligible within the context of the story. Mental state ending: The plot is organized in such a way that the final element of the story is a character's feeling ("she was sad"), expressive behavior ("she cried"), or state of being ("then she was alone"). Ineffective actor: The final element of the story is phrased in passive voice or impersonal construction. It is explicitly stated that characters are helpless or that action is futile and pointless.

The scores for each picture (theoretical range = -3 to +2) are summed to create a single total score (theoretical range = -18 to +12). In this sample, the mean self-definition score was -.12 and the standard deviation was 3.40.

The interview transcripts were coded by a student assistant trained by me on pretest transcripts.⁵ This coder was blind to the identities and scores of the interviewees and was unaware of the nature and hypotheses of the study. The coder and I achieved intercoder reliability (category agreement) above .90 for all categories scored. The categories for coding the interviews can be summarized as follows:⁶

Overall ratings of the interview. The coder was instructed to perform an overall rating of the total interview for the following variables: (a) Clarity of formulation of the problem—Code the problem from 0 (quite unclear) to 2 (quite clear). (b) Clarity of explanation of causes of the problem—Again, rated clarity of the explanation of the causes of the problem on a 3 point scale from "unclear" to "clear." (c) Locus of the problem. Indicate whether the problem is described by the person as being located either within herself or outside herself. (d) Locus of the solution to the problem—Indicate whether the person sees the solution to the problem as located within her own control or outside her own control. (e) Content of problem—Indicate which of the following areas best describes the

Table 1
Self-Definition and the Content of the Problem

Content of problem	<i>n</i>	<i>M</i>	<i>SD</i>
Work or school ^a	14	.57	2.87
Relationships	15	-.13 ^b	2.67
Psychological	8	-.88	3.23
Death or illness	14	-.36	4.70

Note. $F = .51$, *ns*.

^a Including housework, and so forth.

^b Since some of the self-definition categories are scored -1, negative total or average scores are possible.

content of the problem described: (i) work or school (including work at home, if issue is *work*), (ii) interpersonal relationships, (iii) emotional and psychological problems (depression, alcoholism, etc.), or (iv) death and illness.

Type of response to problems. The coder was instructed to code the problem solutions as either: (a) No action—The subject indicated that she did nothing to solve the problem at all (e.g., "I did nothing," "I just waited," "Nothing could be done," etc.) or waited for the problem to be solved by some outside agent (e.g., "My husband got transferred so it got better."). (b) Noninstrumental action—The subject indicated that the solution to the problem was "distraction." That is, the subject's solution to the problem was to immerse herself in some *other* unrelated activity that helped her ignore the problem (e.g., "I was afraid of what would happen if I told him how I felt, so I joined a lot of clubs to keep myself busy," "I took up bridge"). (c) Instrumental action—The subject indicated that she solved the problem by initiating some action or set of actions (e.g., "I went back to school," "I told him how I felt," "I got a job," etc.) that was relevant to the problem at hand.

Results

First, as can be seen in Table 1, there were no differences in self-definition scores of women reporting problems in each of the content areas (work or school, other relationships, etc.). Any differences in the interpretations of self-defining and socially defined women's problems cannot reasonably be attributed to differences in the types of problems that they faced.

⁴ A detailed scoring manual, with illustrative examples and sets of practice stories with associated scoring, can be obtained by writing to the author.

⁵ The author is grateful to Gwen Arthur for this assistance.

⁶ Detailed instructions for coding can be obtained by writing to the author.

The first substantive hypotheses tested involved subjects' interpretation of their problems. As can be seen in Table 2, these hypotheses were supported. Because subjects whose problem involved death or illness might be less likely to hold alternative interpretations of their problem, these subjects were excluded from these analyses. However, results that included these subjects were substantially the same as those reported here. In both analyses, women who perceived their problem as located outside themselves and perceived the solution to their problem as located within themselves were significantly higher in self-definition. Similarly, women who cited more causes of their problem scored higher (though not significantly) in self-definition. Scores were also significantly correlated with the clarity of articulation of the problem ($r = .64$, $p < .001$) and clarity of explanation of the causes of the problem ($r = .73$, $p < .001$). Finally, women who presented their problems as involving a wider social context than the

Table 3

Self-Definition and Alternative Responses to Problems

Response	<i>n</i>	<i>M</i>	<i>SD</i>
Instrumental action	16	1.56*	3.71
Noninstrumental action	10	-2.10	3.14
No action	25	-.44	2.86

Note. $F(2, 48) = 4.25$, $p < .05$.

* This mean is significantly different from the means of those taking noninstrumental actions ($t = 2.59$, $p < .01$, one-tailed) and of those taking no action ($t = 1.89$, $p < .05$, one-tailed).

nuclear family were significantly more self-defining than women who did not.

The final hypothesis tested here was that self-definition would be associated with instrumental rather than noninstrumental or passive responses. As can be seen in Table 3, this hypothesis was also supported. Women who made instrumental responses were significantly higher in self-definition than women who made noninstrumental responses or no responses at all.

Table 2
Self-Definition and Interpretations of the Problem

Item	<i>n</i>	<i>M</i>	<i>t</i>
Locus of the problem*			
Within self	18	-1.67	4.10***
Outside self	19	1.53	
Locus of the solution*			
Within self	17	1.35	2.97**
Outside self	20	-1.20	
No. causes of problem			
More than 3	25	.67	1.57*
Fewer than 3	26	-.85	
Context of the problem			
Broad (outside home)	32	1.02	4.15***
Narrow (only home)	19	-2.00	

Note. All p values are one-tailed.

* Problems classified as "death and illness" were excluded in these cases, since by definition they involved events both outside the self and outside the control of the self. Problems open to alternative interpretations (work, relationships, psychological) were included in the analysis.

* $p < .07$.

** $p < .005$.

*** $p \leq .001$.

Discussion

It is clear that self-defining and socially defined women do indeed interpret the problems that they face differently. As predicted, self-definition in women is associated with seeing one's problems as external but the solutions as within one's grasp (or internal). It is also related to seeing the causes of problems as multiple, and to explaining those causes clearly. Finally, self-definition is high among those who see their problems as part of a relatively wide social context, rather than in a narrowly domestic one. The world view expressed in these women's TAT stories is, then, reflected in the way they construe their own problems, 14 years later. Indeed, it can be said that the categories scored in the interviews are closely related to those scored in the TAT stories, and that therefore the relationships found here might better be interpreted as reliability estimates for self-definition rather than as genuine relationships between independent variables. Even if this argument is accepted, the consequent estimate of the stability of self-definition (.60-.70) over 14 years is itself remarkable enough. However,

the original measure was based on fantasy stories told about neutral pictures; the current interview coding was based on descriptions of actual problem situations and the methods used to handle those situations. Thus, on grounds of temporal separation, different base data, and different scoring systems, the two measures seem to be sufficiently different to justify the inference that self-definition (a dispositional world view) predisposes certain kinds of interpretations of problematic situations.

Finally, it seems reasonable to suspect that self-definition is also related to the actual problem-solving responses that the subjects reported having made. Though there is no independent evidence that these responses were actually made by the subjects, I have no reason to believe that the subjects distorted their responses, particularly since so few reported instrumental responses (31%), which are presumably the more socially desirable ones.

In summary, it is clear that self-definition and social definition are two implicit belief systems about the nature of external reality and one's own position within it with implications for an individual's interpretation of specific personal problem situations and for behavioral responses to those situations.

The implications of these findings should be clear. First, theoretically, the confirmation of a close link between individuals' general implicit belief system and their specific interpretations of and responses to particular problems indicates that coping responses, however "irrational," when viewed by an outside observer, may "make sense," given the world view of the subject. Moreover, this study indicates that it may be useful to examine both general stable belief systems (not only specific, situationally based attributions) and belief systems that are not necessarily conscious but are instead implicit in one's habits of thinking and organizing experience. Second, and more practically, if we wish to change ineffective coping strategies, we may find that it is sometimes useful to

intervene at the level of the global world view rather than at the particular response level. In addition, it is clear that as a construct, the predictive and theoretical utility of self-definition extends beyond the domain of career orientation and career activity. Finally, one implication of this study is that individuals' characteristic world views may often influence their particular judgments in stress situations. Further research examining the role of self-definition, and related dispositional variables, in determining alternative appraisals of threats in the laboratory (see, e.g., Lazarus, 1968) or alternative cognitive "sets" in problem-solving situations (see Scott & Howard, 1970) seems appropriate, given the strong association found in this study between self-definition and individuals' interpretations of their own personal problems.

References

- de Charms, R. *Personal causation: The internal affective determinants of behavior*. New York: Academic Press, 1968.
- Kelley, H. H. *Attribution in social interaction*. Morristown, N.J.: General Learning Press, 1972.
- Lazarus, R. S. Emotions and adaptation: Conceptual and empirical relations. In W. Arnold (Ed.), *Nebraska Symposium on Motivation* (Vol. 16). Lincoln: University of Nebraska Press, 1968.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Scott, R., & Howard, A. Models of stress. In S. Levine & N. A. Scotch (Eds.), *Social stress*. Chicago, Ill.: Aldine, 1970.
- Seligman, M. E. P. *Helplessness: On depression, development and death*. San Francisco: Freeman, 1975.
- Stewart, A. J. *Longitudinal prediction from personality to life outcomes among college-educated women*. Unpublished doctoral dissertation, Harvard University, 1975.
- Stewart, A. J., & Winter, D. G. Self-definition and social definition in women. *Journal of Personality*, 1974, 42, 238-259.
- Winter, D. G., Stewart, A. J., & McClelland, D. C. Husband's motives and wife's career level. *Journal of Personality and Social Psychology*, 1977, 35, 159-166.

Received August 19, 1977 ■

Self Versus Ideal Self: A Comparison of Five Adjective Check List Indices

Harrison G. Gough

Institute of Personality Assessment and Research
University of California, Berkeley

Renato Lazzari and Mario Fioravanti

University of Rome

Five self versus ideal-self measures were defined on the 300-item Adjective Check List. Overall congruence was indexed by the phi coefficient for items and by the sum of the absolute differences on standard scores for the 24 scales (D-T). The absolute differences on the 24 scales were also correlated and factored in two samples: 100 American Air Force officers and 95 Italian young men applying for a national precollege military training school. Three factors common to both samples were identified. D-1, D-2, and D-3 measures of dissimilarity were obtained by summing the absolute differences on just those scales assignable to each of the three factors. Analyses of observers' ratings in the sample of American officers revealed phi, D-T, and D-1 to be indicative of superior personal and social adjustment and D-2 to suggest goal-oriented efficiency and diligence. D-3, in contrast, had rather unfavorable connotations. It is concluded that internal components of self-ideal congruence have differential implications that overall measures will obscure or even fail to detect.

A compelling aspect of the phenomenology of human experience is the sense of individuality, locus, wishes, fears, capacities, and continuity that behavioral scientists refer to as the *self*. By means of introspection and the observation of what others say and do, each person gradually evolves a notion of who he is and how he resembles and differs from other people. Various writers have attended to the way in which this self-concept or sense of identity may reflect social consensus (e.g., the "looking-glass self" of Cooley, 1902), the consequences of role-taking activity in childhood and later (Mead, 1934), and a growing awareness of the differences between what one wants and what the world is prepared to give (James, 1890).

The notion of an ideal self—what one would like or feels constrained to be—can also be posited. McDougall (1932), for example,

described a "self-regarding sentiment" and stressed the active process of comparison that goes on between the actual and the ideal self. Socialization, according to McDougall, is in part the resultant of a continuously more effective and encompassing reconciliation of these two selves. Allport (1961), in his concept of the *proprium*, sought to bring together seven facets of selfhood, including the bodily self, continuity, self-esteem, extension, imagination, rational coping, and goal-directed or propiety striving. The degree to which these facets of the self are harmoniously integrated will determine the degree to which an individual becomes what he or she is capable of becoming.

In this article we shall be dealing primarily with contrasts between the real and ideal selves, or aspects of the self, and with ways in which their congruence or discrepancies can be measured. There have, of course, been many attempts to assess these differences (Wylie, 1974), making use of methods such as the *Q* sort (Butler, 1968; Butler & Haigh, 1954), the semantic differential (Pervin &

Requests for reprints should be sent to Harrison G. Gough, Institute of Personality Assessment and Research, University of California, Berkeley, California 94720.

Lilly, 1967), the Interpersonal Check List (Leary, 1957), and the Minnesota Multiphasic Personality Inventory (Rosen, 1956). For the most part, these analyses have used a single score or index, based either on the responses to individual items under two conditions of administration, or on differences between scores on the scales of the instrument. The possibility that internal components or dimensions of similarity may be identified should also be considered. An overall index of congruence may mask or obscure important internal facets of similarity having different diagnostic implications.

Method

Subjects

Two samples were studied. The first consisted of 95 Italian males (M age = 15.5) who were applying for a national educational program sponsored by the Department of Defense. In educational background these applicants have the equivalent of 10 or 11 years of schooling in the American system. Successful applicants receive free education, food, and lodging, and incidental financial aid, much as do cadets or midshipmen in American military academies. A student in the Italian program may, on graduation, go on to one of the Italian military or naval academies, and the school, in fact, is viewed as a preparatory center for such training. The 95 applicants included in this study had come from all parts of the country to a testing center in Naples.

The other sample was composed of 100 male Air Force officers seen at the Institute of Personality Assessment and Research in Berkeley (MacKinnon, 1958). All held the rank of captain. Their mean age was 33.6, mean years of education 13.7, and mean years of service 11. Each officer spent 3 days at the Institute, participating in interviews, leaderless group discussions, testing sessions, laboratory exercises, and informal activity in which his behavior was observed and rated.

Tests

The principal testing device for this study was the Adjective Check List (ACL; Gough & Heilbrun, 1965). The ACL consists of 300 adjectives or adjectival phrases, from which the respondent selects those believed to be self-descriptive. The number of items checked is unspecified, as variation in endorsement is itself viewed as a personality variable. In addition to a score on number of items checked, there are 23 other scales, on all of which the standard scores are adjusted according to the number of items that are endorsed; this adjustment removes the influence of acquiescence from the 23 measures. Among the scales are those for self-confidence, self-control, achievement, dominance,

aggression, and deference. Testing time varies from 5 to 10 minutes. Italian norms for the ACL (Gough, 1968) were used for standardizing the scores of the 95 Italian males in this study.

The ACL was administered twice, once with normal instructions and the second time with a request to describe one's own ideal, defined as "the kind of person you would like to be—your personal ideal."

For the American sample, each officer was described on the ACL by 10 staff observers from the Institute. Tallies were made of the number of times each word had been checked, and these sums were used as descriptive scores. For example, if all 10 observers described an officer as "boastful," then his score on this item would be 10. If only 2 observers described him as "cooperative," then his score on this item would be 2. In this way 300 descriptive scores were obtained for each of the men.

Five ACL indices

To indicate the correspondence between self and ideal-self descriptions, five measures were used. The first of these was the phi coefficient for the adjectives checked and not checked in the two conditions. This coefficient reflects the degree of overall agreement between the descriptions of self and ideal self. The second index, called *D-T*, was also an indicator of overall agreement. It was defined as the sum of the absolute differences between the 24 standard scores on the ACL profiles for self and ideal self.

To develop indices of internal components, several methods were tried. One was to factor analyze the two 24×24 matrices of standard scores for the ideal self and then to compute difference scores separately for the scales included in each ideal-self factor. In each sample there were five ideal-self factors with eigenvalues greater than 1. Another approach was by way of the canonical correlations between the self and ideal-self scales in the two samples; the self-descriptive weights could then be used to compute estimates of the ideal vectors. These analyses identified two canonical variates in each sample. A third method was to factor analyze the 24 difference scores, based on the absolute discrepancies between the self and ideal-self scores for each of the scales in the checklist. Being based directly on the similarity of scores under the two conditions of testing, this third method appeared to be conceptually closest to the phenomenological notion of congruence between self and ideal self, and it was the one selected for subsequent analyses.

In the factor analyses for both Italian and American samples, there were four factors with eigenvalues equal to or greater than 1. After varimax rotation, these factors were compared for equivalence. The American factors 1, 2, and 3 were correlated .47, .58, and .68 with the Italian factors 2, 1, and 3. The fourth factors from each analysis, however, were ambiguously related and were therefore dropped from further consideration.

On the basis of these factor analyses, three difference scores were computed, each based on the absolute sum of the standard score discrepancies for the scales assigned to each factor. Specifically, *D-1* was based

Table 1
Comparison of Real Versus Ideal Self Adjective Check List Protocols for the American and Italian Samples

ACL scale	Air Force officers ^a				Italian students ^b			
	Real		Ideal		Real		Ideal	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Number Checked	46.94	8.43	48.49	7.91	50.04	8.19	51.28	11.23
Defensiveness	57.49	9.41	60.65**	7.35	54.04	7.34	55.89	8.64
Number Favorable	52.98	9.24	59.85**	8.52	55.01	7.46	57.88*	9.69
Number Unfavorable	47.20**	9.09	41.04	2.04	44.72**	5.96	41.86	5.67
Self-confidence	52.73	10.66	59.16**	7.92	53.34	8.89	58.21**	8.36
Self-control	54.69	9.75	60.28**	5.75	50.95	8.30	55.88**	7.10
Libility	46.06	10.63	49.89**	7.83	51.39	10.43	52.14	9.01
Personal Adjustment	54.26	9.51	58.87**	6.33	53.67	7.75	53.93	7.67
Achievement	58.09	10.03	62.70**	7.10	54.82	8.13	58.58**	7.81
Dominance	56.08	9.47	59.92**	6.30	56.54	7.46	60.72**	7.90
Endurance	58.48	9.79	63.90**	5.65	54.00	9.26	56.85**	6.40
Order	57.28	9.06	60.28**	5.96	52.81	9.87	55.63**	6.44
Intracception	52.79	10.19	57.78**	8.62	52.46	7.21	53.16	8.33
Nurturance	53.41	9.33	55.84**	6.34	52.38	6.24	52.27	8.46
Affiliation	54.94	8.24	56.58	7.53	54.78	7.45	52.56	9.40
Heterosexuality	47.32	8.58	52.62**	7.64	50.11	8.63	53.32**	10.10
Exhibition	48.07	9.76	49.91	5.50	50.78	7.62	52.28	6.80
Autonomy	46.83*	9.36	44.46	5.40	49.20	6.80	48.03	7.89
Aggression	47.13**	9.17	44.76	6.06	47.44	6.62	47.33	7.33
Change	44.67	8.08	43.77	6.88	49.78**	9.26	45.88	7.32
Succorance	43.88**	7.78	39.42	4.37	43.60**	7.65	38.59	6.85
Abasement	45.55*	7.96	43.44	4.59	45.19**	7.34	41.37	6.93
Deference	50.36	10.04	49.11	6.87	48.42*	6.70	46.02	8.58
Counseling Readiness	46.79**	8.53	39.23	6.36	44.42**	8.22	41.63	7.51

^a *n* = 100.

^b *n* = 95.

* *p* < .05.

** *p* < .01.

on the scales from the American first factor and Italian second; it was defined as the sum of the absolute discrepancies on the scales for Number of Items Checked, Number of Favorable Items Checked, Personal Adjustment, Intracception, Nurturance, Affiliation, and Aggression. D-2 included the scales from the American second factor and Italian first; it was defined as the sum of the absolute discrepancies on the scales for Defensiveness, Self-confidence, Achievement, Dominance, and Endurance. D-3 was based on the two third factors, summing the absolute discrepancies on the scales for Libility, Exhibition, Autonomy, Change, Abasement, and Deference.

Results

Table 1 presents standard score means and sigmas for the self and ideal-self protocols of the American and Italian samples. As would

be expected, scores were generally higher under ideal-self instructions for scales measuring favorable or positive variables such as self-confidence and achievement and lower for scales measuring less desirable attributes such as the number of unfavorable words and abasement.

There were 12 differences that were in the same direction and that also yielded statistically significant ($p \leq .05$) *t* tests in both samples. If attention is paid only to the direction of difference, agreement was found in 22 of the 24 comparisons. A binomial test for this consistency gave a *z* value of 3.60, significant well beyond the .01 level of confidence. It can be concluded that there is appreciable correspondence in the two samples

Table 2

Means, Standard Deviations, and Intercorrelations of the Five Indices of Self Versus Ideal-Self Congruence in the American and Italian Samples

Index	Intercorrelation ^a					American ^b		Italian ^c	
	1	2	3	4	5	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
1. Phi	—	-.79	-.71	-.64	-.43	.50	.21	.42	.25
2. D-T	-.71	—	.82	.80	.65	190.45	87.47	183.80	77.97
3. D-1	-.68	.74	—	.50	.33	54.78	30.63	52.39	30.57
4. D-2	-.40	.78	.37	—	.33	41.16	28.05	42.48	24.29
5. D-3	-.51	.70	.28	.45	—	47.96	25.35	43.54	24.84

Note. D-T = the sum of the absolute differences between the 24 standard scores on the Adjective Check List profiles for self and ideal self; D-1 = the sum of the absolute differences on seven scales; D-2 = the sum of the absolute differences on five scales; and D-3 = the sum of the absolute differences on six scales. See text for specification of these scales.

^a Coefficients for the American sample are above the diagonal; coefficients for the Italian sample are below the diagonal.

^b *n* = 100.

^c *n* = 95.

in regard to the effect of the "ideal self" instructional set.

The five indices of agreement or disagreement between self and ideal-self description were next computed for the men in each sample. The means and standard deviations for each index are given in Table 2, along with the intercorrelations among the five measures of congruence.

The two overall indicators of similarity (phi and D-T) correlated $-.79$ in the American and $-.71$ in the Italian sample. These coefficients are in the expected direction, as phi reflects agreement and D-T reflects disagreement between the two conditions of testing. Although both coefficients are high, they are less than perfect and suggest that the relative degree of correspondence between the self and ideal-self descriptions of an individual will depend in part on the measure used to assess that correspondence.

The three remaining indices are each based on nonoverlapping subsets of scales from the ACL. For the American sample the median correlation among the D-1, D-2, and D-3 indices was .33, and for the Italian sample the median was .37. These coefficients are low enough to suggest that the three measures may be assessing different facets of the self versus ideal-self correlation.

The differences between the means of the American and Italian samples on the five

indices were all statistically insignificant, suggesting that these values may be taken as reliable benchmarks.

To examine the differences in diagnostic implications among the five measures, data were taken from the study of Air Force officers. As was mentioned earlier, each officer was described on the ACL by 10 staff observers. The number of times that each item was checked by these panels was taken as the descriptive score on the adjective; these 300 scores were then correlated with the five indices. To simplify the problem of interpreting findings, only the six adjectives with highest positive correlations and the six with largest negative correlations were selected for each of the five indices.

For the phi coefficient of similarity between self and ideal self, the six descriptions with highest positive correlations were cooperative (.28), adaptable (.26), planful (.26), outgoing (.25), efficient (.24), and thorough (.24). These coefficients are all significant beyond the .05 level of probability, and taken together they suggest a quite favorable picture of the individual whose self and ideal-self descriptions are similar. The six terms with largest negative correlations were slow ($-.25$), foolish ($-.24$), awkward ($-.22$), confused ($-.22$), unrealistic ($-.21$), and unfriendly ($-.18$).

For the D-T index of dissimilarity, we shall report first those adjectives having negative

correlations, as they are related to closer correspondence between self and ideal-self descriptions. The six with largest coefficients were thorough ($-.24$), planful ($-.20$), adaptable ($-.19$), cooperative ($-.19$), loyal ($-.18$), and reasonable ($-.18$). The cluster is distinctly favorable, but the degree of relationship is weaker than that for the phi coefficient. The six terms with largest positive values were headstrong (.26), opinionated (.25), boastful (.24), tactless (.24), arrogant (.23), and individualistic (.22).

For the D-1 index of dissimilarity, based on 7 of the 24 scales in the ACL, the six terms with largest negative correlations (and therefore associated with greater congruence of self and ideal-self descriptions) were thorough ($-.26$), cooperative ($-.25$), loyal ($-.25$), reasonable ($-.24$), considerate ($-.23$), and stable ($-.22$). The six terms with the largest positive correlations were boastful (.36), headstrong (.35), tactless (.33), opinionated (.32), arrogant (.31), and self-centered (.30).

Although there are differences in the particular words cited, the general tenor of the characterizations of officers showing greater self-ideal congruence on these three measures is favorable. Officers whose self-descriptions more closely approximate their ideal selves are viewed as more cooperative and less headstrong.

If the diagnostic implications of the three indices are more or less the same, which of the three measures can be recommended? One answer to this query can be based on the time needed for computation. For hand calculation, D-1 is easier to use than either D-T or phi, assuming that the two ACL protocols have been scored and profiled. Another answer can come from the certainty of the diagnostic implications. The median coefficient among the 12 given for each index can provide a crude measure of this certainty. For the phi index the median correlation was .245, for D-T it was .235, and for D-1 it was .280. The differences are slight, but they favor D-1.

Correspondence between self and ideal self has typically been interpreted as a measure of personal adjustment and stability. In the studies of psychotherapeutic outcome reported by Butler and Haigh (1954), an increase in the correlation between self and ideal-self *Q* sorts was accepted as a sign of improvement.

Wylie (1974), after reviewing a large number of studies, concluded that congruence between self and ideal-self description was generally taken as an indicator of self-acceptance and personal adjustment. The descriptive correlation found in our Air Force sample for phi, D-T, and D-1 gives some support to these inferences. A cautionary note should be sounded in regard to "oversatisfaction with self" (Block & Thomas, 1955). It is possible that too much congruence between self and ideal-self description could reflect insensitivity to personal problems, defensiveness, narcissism, and other undesirable attributes. A phi coefficient of .99 between self and ideal-self ACL protocols would not be viewed very favorably even by the most ardent advocate of the index. In our two samples, it should be remarked, exceedingly high phi coefficients were not encountered. For the Air Force sample, the highest coefficient was .85, and in the Italian sample the highest phi was .84.

In the Air Force sample, several ratings of personal soundness and social adjustment were available, contributed by the staff members who had interviewed and studied the officer. The rating for personal soundness correlated .13 with the phi index of correspondence between self and ideal self, $-.11$ with D-T, and $-.23$ with D-1. If an index of self versus ideal self is intended to carry a diagnostic implication of personal soundness or stability, it would appear from these three coefficients that the implication is strongest for the D-1 index. Another indication of adjustment came from a preliminary form of the Block *Q* sort (Block, 1961). An item in that form stated "Gets along in the world as it is; is socially appropriate in his behavior; keeps out of trouble." The mean staff placement of this item correlated .21 with phi, $-.14$ with D-T, and $-.26$ with D-1. Once again the advantage lies with the D-1 index. A final example can be taken from the Block item "is socially perceptive, responsive to interpersonal nuances." Correlations with mean staff placement of this item were .23 for phi, $-.20$ for D-T, and $-.28$ for D-1.

The second internal measure was D-2, based on the scales for Defensiveness, Self-confidence, Achievement, Dominance, and Endurance. When D-2 was correlated with the descrip-

tions, the six with the largest negative coefficients (i.e., checked more often about officers whose two protocols were more similar) were playful ($-.25$), efficient ($-.22$), adaptable ($-.21$), ambitious ($-.21$), sharp-witted ($-.21$), and thorough ($-.21$). The six descriptions with the largest positive correlations were slow (.28), shy (.26), awkward (.24), commonplace (.22), unaffected (.22), and slipshod (.21). Once again, officers whose two reports were more congruent were described more favorably, and those whose reports were less congruent were described less favorably.

In this regard D-2 resembles phi, D-T, and D-1. What, if any, are the differences? One difference is that D-2 carried no implications for the Q-sort item "gets along well in the world as it is," for which the coefficient was zero. For the rating of personal soundness, the coefficient for D-2 was $-.01$. Thus, although D-2 has implications for planfulness and efficiency, it does not have the implications that D-1 revealed for personal soundness and everyday social adjustment. On the Q-sort item "lacks confidence in own ability," the correlations for phi, D-T, D-1, and D-2 were $-.10$, $.02$, $-.10$, and $.19$, respectively. The sign of the coefficient for D-2 indicates that officers whose two descriptions were more discrepant were characterized by this lack. D-2, it appears, is reflective of a kind of problem-solving effectiveness but not of personal adjustment or soundness.

The third internal measure of congruence, D-3, may now be considered. The six adjectival descriptions correlating most strongly with similarity on D-3 were greedy ($-.20$), unselfish ($-.19$), fearful ($-.18$), self-punishing ($-.17$), ingenious ($-.16$), and changeable ($-.15$). The correlations for these six terms are low, and all save one are less than the coefficient of $.195$ necessary for significance at the $.05$ level of probability. It should be observed, nevertheless, that three of the six are clearly unfavorable in implication and that only two (unselfish and ingenious) are clearly favorable. D-3 therefore represents a distinct departure from the favorable implications of similarity between self and ideal found for phi, D-T, D-1, and D-2.

The six descriptions correlating most strongly with dissimilarity of the two protocols

on D-3 were daring (.25), tough (.24), hard-hearted (.23) attractive (.19), opinionated (.19), and strong (.19). Although these six relationships are of borderline statistical significance, they do cohere in a meaningful manner and suggest an individual possessing vivid and to some extent invasive social characteristics.

There were two ratings for which the correlations with D-3 were larger than for the other four indices. One of these was that for "ability to obtain sexual gratification," obtained from the psychiatric life history interview. The coefficients were $.05$ for phi, $.10$ for D-T, $.08$ for D-1, $.08$ for D-2, and $.18$ for D-3. The other was for the staff rating of apparent health and vitality, for which the coefficients were $.07$ for phi, $.12$ for D-T, $.09$ for D-1, $.07$ for D-2, and $.16$ for D-3. These are very low correlations, but they are of interest in that they show officers who are more discrepant on the D-3 index to possess more zest and vigor.

Discussion

One goal of our analysis was to determine whether or not internal components within the self versus ideal-self context might have implications different from each other and different from a global or overall index of congruence. We also wished to compare two methods of assessing overall congruence, one based on the scales of the ACL and the other taken directly from its items. A third aim was to enhance the generalizability and reliability of the measures by drawing on cross-cultural data in their derivation. In order to identify internal themes or components, factor analyses of difference scores for the 24 scales of the ACL on self and ideal-self protocols were conducted in American and Italian samples. Three of the four factors extracted in the two samples were compatible and were used to define three internal themes.

The two overall indices defined by the phi coefficient and a discrepancy score (D-T) based on all 24 scales were highly correlated with each other and gave rise to similar patterns of descriptive implications when correlated with adjectival descriptions and ratings in the American sample of 100 Air Force officers. The first internal component

(D-1) also correlated significantly with these two overall indices and produced a similar pattern of diagnostic implications. In general, the pattern for all three indices was indicative of positive personal adjustment, stability, and facility in coping with the world as it is. The level of these diagnostic implications was generally highest for the D-1 index. On the ACL, it follows, D-1 would be the best of the three measures of congruence between self and ideal self, if congruence is to be used as an indicator of personal adjustment and self-acceptance.

The other two internal indices differed somewhat from D-1 in their diagnostic implications. D-2 put more stress on personal efficiency, ambition, and work-related attributes and less on personal soundness and everyday social adjustment. D-3 revealed officers whose two protocols were more congruent to be more fearful, changeable, and self-punishing, whereas those whose protocols were more discrepant were more attractive, more daring, and more capable of seeking and attaining their goals in the sexual sphere.

Because of these differences in the diagnostic implications of the three internal indices, it is apparent that an overall index such as D-T must mask or conceal these variations. For example, the relatively unfavorable implications of congruence on the six scales included in the D-3 index are masked when a total discrepancy score based on all 24 scales is used. Differentiated measurement of similarity between self and ideal-self description is therefore to be preferred. The three ACL indices presented in this article may serve as examples of such an analytic perspective, even if future study finds other internal measures of congruence to be preferable. An obvious cautionary note, in this regard, is that the three components of congruence reported above were derived from samples of males only. Analyses of samples of females must be carried out to determine whether these or other indices of congruence will be most valid. Other ways of defining the ideal self could also be considered, for example, "the person my parents want me to be," "the person I would like to be in 10 years," and "myself, when I am at my best." The essential findings in our analysis, we conclude, are (a) that there are facets or aspects of

the congruence between self and ideal and (b) that these facets have different diagnostic implications. It follows that psychometric methods for assessing self-ideal congruence should include measurement of internal components as well as of overall or general correspondence.

References

- Allport, G. W. *Pattern and growth in personality*. New York: Holt, Rinehart & Winston, 1961.
- Block, J. *The Q-sort method in personality assessment and psychiatric research*. Springfield, Ill.: Charles C Thomas, 1961.
- Block, J., & Thomas, H. Is satisfaction with self a measure of adjustment? *Journal of Abnormal and Social Psychology*, 1955, 51, 254-259.
- Butler, J. M. Self-ideal congruence in psychotherapy. *Psychotherapy*, 1968, 5, 13-17.
- Butler, J. M., & Haigh, G. V. Changes in the relation between self-concepts and ideal concepts consequent upon client-centered counseling. In C. R. Rogers & R. F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago: University of Chicago Press, 1954.
- Cooley, C. H. *Human nature and the social order*. New York: Scribners', 1902.
- Gough, H. G. Lo sviluppo di una taratura italiana per l'Adjective Check List (The development of Italian norms for the Adjective Check List). *Bollettino di Psicologia Applicata*, 1968, n. 88-90, 23-31.
- Gough, H. G., & Heilbrun, A. B., Jr. *The Adjective Check List Manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1965.
- James, W. *Principles of psychology* (Vol. 1). New York: Holt, 1890.
- Leary, T. F. *Interpersonal diagnosis of personality*. New York: Ronald Press, 1957.
- MacKinnon, D. W., Crutchfield, R. S., Barron, F., Block, J., Gough, H. G., & Harris, R. E. An assessment study of Air Force officers: Part I. Design of the study and description of the variables. (Wright Air Development Center Technical Report No. 91 (I)). Lackland Air Force Base, Tex.: Personnel Laboratory, 1958.
- McDougall, W. *The energies of men*. London: Methuen, 1932.
- Mead, G. H. *Mind, self, and society*. Chicago: University of Chicago Press, 1934.
- Pervin, L. A., & Lilly, R. S. Social desirability and self-ideal self on the semantic differential. *Educational and Psychological Measurement*, 1967, 27, 845-853.
- Rosen, E. Self-appraisal and perceived desirability of MMPI personality traits. *Journal of Counseling Psychology*, 1956, 3, 44-51.
- Wylie, R. C. *The self-concept* (Rev. ed.). Lincoln: University of Nebraska Press, 1974.

Skill Training With Alcoholics

Edmund F. Chaney and Michael R. O'Leary
Veterans Administration Hospital, Seattle, and
Department of Psychiatry and Behavioral Sciences
University of Washington

G. Alan Marlatt
Department of Psychology, University of Washington

This study evaluated a short-term skill-training intervention that taught male alcoholics generation of appropriate behaviors in problematic situations. Forty alcoholics engaged in inpatient treatment were divided into three groups—a skill-training group, a discussion group, and a no-additional-treatment control group. A verbal role-playing measure of responses to situations associated with drinking behavior and relapse showed significant performance improvement of the training group as compared to the control groups. A 1-year posttreatment follow-up indicated that skill training decreased the duration and severity of relapse episodes. Behavior on the situational role-playing task predicted post-treatment adjustment. While pointing out limitations of skill training as implemented, results suggest its utility as one component of a multimodal behavioral approach to relapse in problem drinking and other problem areas such as drug addiction, smoking, obesity, and crime.

A social-learning approach to problem drinking (Bandura, 1969) suggests that in addition to the psychophysiological effects of alcohol, other factors such as cultural and subgroup mores, learning experiences within the family, peer modeling, instrumental functions, and expectancies are relevant to problem drinking. Thus, each individual's drinking behavior is likely to have multiple determinants and would require treatment incorpo-

rating a variety of techniques—a broad spectrum approach (Hamburg, 1975; Nathan, 1976).

One of the treatment strategies of broad spectrum approaches is to provide the problem drinker with the means of achieving reinforcement other than through the consumption of alcohol. This strategy involves identifying both the discriminative stimuli for drinking and the reinforcing consequences, and then making the occurrence of other more satisfying behaviors more likely through a variety of training techniques. Sobell and Sobell (1973), for instance, focused on the drinking behavior of their subjects. Subjects in the experimental group with a controlled-drinking goal were taught skills that theoretically would allow them to not drink or to drink in an appropriate social manner. Sobell and Sobell suggest that subjects who functioned well after discharge had learned to recognize discriminative stimuli for drinking and to generate alternative responses.

There is little systematic evaluation of training problem drinkers to generate alternative behaviors. Several investigators have

This research was supported by the Research Service, Seattle Veterans Administration Hospital, and by the University of Washington Alcohol and Drug Abuse Institute. This study is based on a dissertation by the first author submitted in partial fulfillment for the PhD degree at the University of Washington.

This study would not have been possible without the cooperation of William Hague, former Chief of the Seattle Veterans Administration Alcohol Treatment Program. Among the many students of behavior who took part in this study, the assistance of Ralph Boynton, Judith Chaney, Dennis Donovan, Dorothy Knapp, Flora Ostrow, Henne Queisser, and Demaris Rohsenow was particularly valuable.

Requests for reprints should be sent to Edmund F. Chaney, Psychology Service, Veterans Administration Hospital, 4435 Beacon Avenue South, Seattle, Washington 98108.

outlined training approaches but have presented little evaluative data (Foy, Miller, Eisler, & O'Toole, 1976; Toomey, 1972). Skill-training techniques only recently have begun to receive systematic evaluation in other patient populations (Arnkoff & Stewart, 1975; Curran, 1977; Finch & Wallace, 1977; Goldsmith & McFall, 1975). Skill training is developing as a combination of methods of teaching the performance of new behaviors and of generating appropriate content of those behaviors. Instruction, modeling, coaching, and role playing, or behavioral rehearsal, appear to be additive components of training (Rich & Schroeder, 1976). Evidence also suggests that training tends to be situation specific (Eisler, Hersen, Miller, & Blanchard, 1975). Thus generalization of training is an important issue, and skill training should include systematic teaching of problem-solving skills so that subjects can analyze novel problematic situations and generate and evaluate adaptive responses (D'Zurilla & Goldfried, 1971).

Problem drinkers report that they use alcohol in a wide variety of stressful situations. Studies of drinking behavior in experimental situations help to confirm that heavy drinkers increase their consumption of alcohol when subjected to interpersonal stress (Allman, Taylor, & Nathan, 1972; Higgins & Marlatt, 1975; Miller, Hersen, Eisler, & Hilsman, 1974). To determine whether self-report and experimental findings are related to continued problem drinking, the characteristics of relapse situations need to be examined. Unfortunately, few investigators have reported situational characteristics of relapses as they happen following treatment.

The present study follows Marlatt's (1978) categorization of relapse situations of male alcoholics following aversion-conditioning therapy. It was found that relapse situations could be assigned reliably to the following types: (a) frustration and inability to express anger (29%); (b) inability to resist social pressure to drink (23%); (c) intrapersonal negative emotional state (10%); (d) inability to resist intrapersonal temptation to drink (21%); and (e) other, or no response (17%). Training situations used in the present study were balanced among the first four categories.

The present study addressed two questions: Does skill training increase problem drinkers' effectiveness in responding to stressful situations? and Does an improvement in psychosocial problem solving have an effect on the subsequent occurrence of problem drinking behavior? The population of problem drinkers considered in this experiment consisted of volunteers in intensive inpatient alcoholism treatment. Subjects were assigned randomly to one of three groups: experimental skill training, discussion control, or no-additional-treatment control. The experimental group used modeling, role-playing, and coaching techniques to work through the problem-solving steps of problem definition and formulation, generation of alternatives, and decision making. Optimal alternatives were rehearsed. The discussion control group talked about the same situations that the experimental group practiced solving, but they did not use role playing, modeling, or coaching. A second control group received no additional training beyond the regular treatment regimen. Training effects were assessed on a verbal role-playing measure of responses to situations associated with drinking behavior and relapse (the Situational Competency Test; SCT). Subjects also were given a structured follow-up interview at 1-month, 3-month, 6-month, and 12-month intervals following discharge. Specific hypotheses tested were: (a) The experimental treatment group would show a pre-post increase in competency in handling problematic situations as measured by the SCT, compared with the competency level of the two control groups, which would not change; and (b) the experimental group would show superior posthospitalization adjustment as compared with the control groups.

Method

Subjects and Treatment Setting

Patients in the Seattle Veteran's Administration Alcoholic Treatment Program (ATP) formed the potential subject pool. All male patients residing in the program 19-26 days and agreeing to stay for the full program were asked to volunteer. Participation was voluntary, and refusal to take part did not jeopardize treatment. All subjects had a primary diagnosis of alcoholism, and none were actively psychotic or organically impaired to the extent of requiring extended

domiciliary care in the opinion of the treatment staff. All subjects continued to participate in the regular treatment program.

Patients admitted to the ATP initially entered an orientation phase in which detoxification was completed and acute medical problems were resolved. After a minimum of 5 days, the patient progressed to the evaluation phase, a 2-week period during which the patient attended six 60-minute small group therapy sessions, videotaped replays of the group sessions, daily therapeutic community groups, and several education-oriented lectures and movies. The program was abstinence oriented, and group therapy, the primary therapeutic modality, examined ongoing relationships and feelings in the patient community with the assumption that understanding and working out relationships with other group members would carry over to important outside relationships, obviating the need for alcohol.

At the end of the 2-week evaluation period, the ATP required patients to make a commitment for an additional 2 weeks of inpatient treatment, a 4-week day program, and 1 year of a weekly aftercare group. Patients who could not make this commitment were referred elsewhere. Treatment activities continued through the day program as described for the evaluation period. Single men were required to live in a supervised halfway house during the day program and for an additional 2 months thereafter. Married men returned home but participated in a weekly couples group for 3 months.

Of the 70 persons asked to participate, 56 consented. Of this group, 6 dropped out of the research before being assigned to a group. An additional 10 patients dropped out during the training phase: 4 were skill-training subjects, 3 were discussion subjects, and 3 were controls. This left 40 subjects who were retested and followed: 15 in the experimental group, 13 discussion controls, and 12 no-additional-treatment controls. Within the limitation of their sequential entry into the program, subjects were assigned to the three groups on a random basis.

Subjects included 2 black men and 1 Native American. Subjects' mean age was 45.6 ($SD = 9.32$). Seventeen men were married; 3 were single; 20 were separated or divorced. The mean number of years of education was 12.3 ($SD = 2.46$). Average total monthly income prior to hospitalization was \$335 ($SD = \470). Mean months employed in current job was 18.6 ($SD = 50.1$). Modal social class was IV (i.e., lower middle) computed from the Hollingshead *Two-Factor Social Position Index* (Hollingshead, Note 1). Seven subjects were court referred; the remainder voluntarily presented themselves for treatment. The average number of years of self-acknowledged problem drinking for the sample was 17.0 ($SD = 10.18$). Subjects reported an average of 3,250 drinks ($SD = 2,450$) consumed during the 6 months prior to admission. Drinks were defined as 1 ounce (29.57 cm³) of 86-proof liquor or its equivalent in alcohol content. Subjects drank an average of 122 days ($SD = 67.1$) during the 6-month period. Eleven subjects reported binge-drinking patterns; 25, steady drinking; and 4, a mixed pattern. The average

number of previous treatment attempts was 1.3 ($SD = 1.27$). Average number of months elapsed since first alcohol treatment attempt was 61 ($SD = 76.1$).

With the exception of the reported number of problems due to drinking, $F(2, 37) = 8.8$, $p < .001$, there were no significant differences among the three subject groups on pretreatment demographic and drinking history measures. The drinking problems measure consisted of a checklist with eight items such as job loss, marital difficulties, and illness. The skill-training group reported more problems prior to treatment than the discussion group ($M = 6.5$, $SD = .9$ vs. $M = 4.4$, $SD = 1.5$, $p < .05$, using Duncan's new multiple-range test). The control group's mean of 5.4 ($SD = 1.5$) was not different from the other groups.

Development of Skill Assessment and Training Procedures

Several sources were used to generate situations likely to be problematic for the present population of excessive drinkers. These included (a) descriptions of relapse situations by Marlatt (1978) gathered in individual interviews with his patients during follow-up; (b) suggestions by treatment personnel of two treatment facilities (Seattle Veteran's Administration Hospital and Mendota State Hospital, Wisconsin) who work with male alcoholics; (c) interviews with alcoholics on the Seattle ATP; and (d) modifications of situations from several inventories designed to assess assertive behavior (Eisler, Miller, & Hersen, 1973; Lawrence, 1970; McFall & Marston, 1970). Eighty situations were assembled and worded to be specific enough to elicit a small number of appropriate courses of action but general enough to be useful as standardized situations for this population. This 80-item inventory (the Situational Difficulty Questionnaire¹) was presented to 40 patients on the ATP who did not take part in any other phase of the study. They were instructed to rank each situation with regard to the difficulty that it would present if encountered in the natural environment. Situations were divided into four categories: (a) frustration and anger; (b) interpersonal temptation; (c) negative emotional state; and (d) intrapersonal temptation. The eight situations judged most difficult within each of the categories were retained for use in the study.

In frustration and anger situations, the person experiences the blocking of a goal-directed activity and/or hostility toward some person or external event. For example:

Before you entered the alcoholism treatment program, your employer, who knew about your drinking problem, said that you could have your job back when you got out of the hospital. When you leave

¹ Copies of the Situational Difficulty Questionnaire, Situational Competency Test, training manual, and follow-up questionnaires for subject and collateral reports are available from the first author.

the program, you find that the company has hired someone to take your place.

In interpersonal temptation situations, the person experiences explicit or implicit pressure by other people to drink. For example:

You are eating at a good restaurant on a special occasion with some friends. The waitress comes over and says, "Drink before dinner?" Everyone else orders one. All eyes seem to be on you.

In negative emotional state situations, the person experiences feelings such as loneliness, depression, boredom, futility, malaise, or nervousness in the absence of clear-cut environmental or interpersonal stimuli. For example:

You get up Saturday morning and realize that you don't have anything planned to do during the day. You sit around for a while, but you begin to feel bored and restless.

In intrapersonal temptation situations, the person experiences a desire or compulsion to drink in the absence of specifically identified external or internal factors. For example:

You have been out of the hospital a couple of months now and haven't taken a single drink. However, you've been wondering how well the treatment really worked, and you get to feeling like taking a drink to test it out.

Assignment of situations into the four categories was assessed independently by two judges (a clinical psychologist and a clinical psychology graduate student) with 94% agreement. Of the 32 situations, 16 were used for training purposes, leaving 16 available for pretreatment and posttreatment testing. No two training and assessment situations were exactly alike.

Pretreatment Measures

Testing situations were assembled into the SCT (see Footnote 1). Situation descriptions were tape recorded, ending with the phrase "What would you do or say?" The subject was instructed to imagine that the situation was actually occurring and to say the words or describe the action that he would use to respond to the situation. Subjects' responses were tape recorded. For the pretest a sample situation was provided to check the subject's understanding of the task. A first version of the SCT was administered to 10 subjects who did not participate in other phases of the research to check that the situation descriptions were clear and that they elicited varied responses.

Four scoring measures, latency, duration, compliance, and specification of new behavior, were chosen. Latency of response was defined as the elapsed time from the termination of the recorded situation to the semantic beginning of the subject's response, excluding disfluencies and hedging. Response duration was the number of words in the response, excluding side comments to the experimenter. Compliance was a dichotomous score indicating whether or not the

subject gave in to the situation without attempting to exert control that would change or influence the course of the situation. Drinking, giving in to a demand, not expressing feelings, and tacitly agreeing to criticism were scored as compliant. Specification of new behavior was also a dichotomous score indicating whether the description of the problem-solving behavior to be performed was given in sufficient detail so that someone else could use the description as a guide to perform the behavior. In other words, a response must have specified one alternative rather than a class of alternatives: "I would come up to the ward" versus "I would get help." This measure was applicable to both interpersonal and intrapersonal situations.

Subjects took the SCT prior to treatment, immediately after, and 3 months following discharge from the hospital. Measures were summed across situations for each subject. All scoring for the four measures across the three administrations was done at the same time by the same rater (a clinical psychology graduate student) who was blind to subject identity and test order. Reliability of the compliance and specification measures was assessed by Pearson correlations of the primary rating of 20 randomly selected protocols of 16 situations, with ratings by an independent rater (also a clinical psychology graduate student; compliance $r = .85$, specification of behavior $r = .82$).

During the evaluation phase of the ATP, the patient received psychological testing including the Shipley Institute of Living Scale (Shipley, 1940), a test of cognitive ability and intellectual impairment. Impairment, expressed by the Conceptual Quotient (CQ), indicates the extent to which the individual's abstract thinking falls short of his or her vocabulary.

After agreeing to participate in the study, all subjects were interviewed using a shortened version of the Drinking Profile (Marlatt, 1976). This structured interview instrument systematically assesses psychosocial functioning and drinking behavior prior to treatment. The Drinking Profile was administered by one of three graduate psychology students. Each had previous clinical experience and was trained by giving at least four supervised administrations of the Drinking Profile. These three experimenters also gave the SCT and conducted most of the follow-up interviews. During the period of the study, they had no involvement in the subjects' treatment, routine or experimental.

Training of Therapists

A training manual (see Footnote 1) was written for the skill-training and discussion treatment procedures. The manual provided an orientation for the therapists, specified introductory statements for the two types of groups, and gave procedural guidelines. It also gave suggestions for working with each of the four situational categories, and for skill training it summarized the problem that each situation presented and gave instructions for evaluating the situational responses. Six therapists were trained to conduct the two groups. The two primary therapists were females: a vocational rehabilitation technician who was experienced in conducting behavioral group therapy and a

clinical psychology graduate student who had completed her internship. Two male clinical psychology graduate students engaged in their internship served as cotherapists during the first half of the study. When they finished their internship, they were replaced by two male recovering alcoholic counselors. The counselors had prior experience leading groups and were both working toward degrees in social work. Therapist training consisted of reading the training manual and background material and role playing the therapy procedures for three 1-hour sessions.

Therapists worked in male-female pairs, and each therapist was involved in an equal number of skill-training and discussion groups. All of the skill-training and discussion sessions were observed by one of the investigators. The observer and therapists met after each group for feedback to minimize drift in technique over time.

Treatment Procedures

Training for the experimental and discussion groups consisted of eight semiweekly 90-minute sessions. Cohorts of 3-5 subjects met with two therapists. Starting dates of skill-training and discussion groups alternated so that if any changes took place in the ATP while the study was underway, the skill-training and discussion groups would be affected equally. As dictated by random assignment, control subjects were distributed individually throughout the duration of the study.

Skill-training group. Skill-training procedures were based primarily on a series of experiments by McFall (McFall & Lillesand, 1971; McFall & Marston, 1970; McFall & Twentyman, 1973), who investigated effective components of assertion training. The content of training was based on D'Zurilla and Goldfried's (1971) stepwise analysis of problem solving. The steps are (a) orientation, (b) definition, (c) generation of alternatives, (d) decision making, and (e) verification.

After giving a general orientation to problem-solving procedures, therapists read a description of a problematic situation. Subjects discussed how they viewed the situation and generated possible ways of responding to it. The therapists pointed out when group members defined the situation differently and what the consequences of different definitions were for problem solving. The probable consequences of the different alternatives proposed by members were discussed, and, if necessary, the therapists proposed alternatives. For interpersonal situations, one therapist chose an alternative, explained the basis of this choice in terms of probable consequences, and then modeled a response, with the cotherapist playing the other person in the situation. For intrapersonal situations one of the therapists engaged in a monologue, explicitly defining the problem, generating alternative solutions, deciding which one would maximize long- and short-term gains and could be performed, and outlining steps to implement the solution. After this initial phase, each group member decided on a particular response and rehearsed it, receiving feedback from the group on the probable consequences of his response. If the therapists and group felt that the response was not likely to solve the

problem that the situation presented, the subject was required to repeat his performance. After each subject had rehearsed, a member summarized the method for generating and evaluating an adequate response to that situation. Two prepared situations were introduced during each session, covering four situations in each of the four categories during training. Subjects also rehearsed one or two situations of their own devising that they felt might be problematic after discharge from the hospital.

In summary, skill-training groups incorporated instruction, modeling, behavioral rehearsal, and coaching, both of actual response behavior and of the cognitive process for generating the response. Subjects were taught how to define the problem that a situation presented by specifying the elements and to generate alternatives and think about the long- and short-term consequences. Finally, the behavior rehearsal phase of training provided practice in carrying out adaptive responses and served as a role-playing form of verification, assessing the adequacy of the problem-solving process.

Discussion group. Discussion group procedures were based on the rationale that problem drinkers do possess the necessary skills to analyze and cope with high-risk situations, but, because of feelings of anger, anxiety, dependency, or depression, they do not effectively use those skills. Discussion procedures focused on eliciting the feelings that may be present in problematic situations and examining subjects' reactions and motivations relevant to similar situations of the past with the logic that self-understanding and more effective coping behavior should result. Thus, in the discussion groups, after giving the orientation, the therapists introduced the same problematic situations as were used in the skill-training group. Then, in a nondirective manner, they encouraged expression and discussion of feelings relevant to the situations.

Control group. A third group of subjects received all assessment measures at the same intervals as other subjects but participated only in regular ATP treatment activities.

Posttreatment Assessment and Follow-Up

Immediately following the eight sessions of treatment, subjects were retested on the SCT. Subjects typically were discharged from the day treatment phase of the program and began weekly aftercare groups the week following completion of the experimental treatment program.

Follow-up procedures were dated from the time the subject left the day treatment phase of the ATP. At 1-month, 3-month, 6-month, and 12-month intervals, subjects were interviewed using a standardized follow-up questionnaire (see Footnote 1), which was designed to provide information compatible with the Drinking Profile. The form assesses occupational stability, living situation, and use of therapeutic supports and provides drinking disposition data by asking about periods and rates of drinking and periods of hospitalization or incarceration. The questionnaire also permits a detailed examination of relapse situation characteristics by

recording place, people present, activity, environmental events, inner thoughts and feelings, and reasons for starting to drink. Subjects had initially consented to have their self-reports verified, and this was done at each follow-up interval through contact with at least one relative, friend, employer, or treatment agency.

At the 3-month interval, the SCT was readministered in person or by telephone in the few cases in which the subject had moved out of the area. In the latter cases, the self-report interview instrument also was administered by telephone. Subjects who had dropped out of treatment and/or moved out of the area often were difficult to locate. Court records, Veteran's Administration files, employers, relatives, friends, and other patients provided leads. In a few cases it was necessary to wait out a period of heavy, sustained drinking before the subject could be interviewed. Throughout follow-up, the subjects' confidentiality was protected.

Results

Situational Competency Test

Heterogeneity. The four variables on which the SCT was scored were intercorrelated for the preadministration, exit, and 3-month postadministration to determine to what extent the derived measures were independent of each other. SCT analyses, which included exit and 3-month posttest data, were based on an n of 37, because 1 discussion group subject's exit data were lost through instrument failure and 2 skill-training subjects' 3-month posttest

Table 1
Situational Competency Test Intercorrelations

Measure	Latency	Duration	Noncompliance
Duration			
Pre	.03		
Exit	.14		
3-month post	.34*		
Noncompliance			
Pre	-.29	.09	
Exit	-.07	.26	
3-month post	-.34*	.08	
Specification			
Pre	-.20	.29	.58***
Exit	.01	.51**	.69***
3-month post	-.05	.20	.36*

Note. $N = 37$.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 2

Effects of Training on the Situational Competency Test: Repeated Measures Analysis of Variance

Measure	Source	df	MS	F
Latency	Group (A)	2	5.6	1.6
	Time (B)	1	.00	<1
	A \times B	2	1.8	1.8
Duration	A	2	1,599.39	5.2**
	B	1	.61	<1
	A \times B	2	326.07	3.4*
Noncompliance	A	2	.05	2.9
	B	1	.00	<1
	A \times B	2	.01	<1
Specification	A	2	.18	8.0**
	B	1	.00	<1
	A \times B	2	.01	<1

Note. The analysis is of exit and 3-month posttest scores residualized on prescores.

* $p < .05$.

** $p < .01$.

data could not be gathered at the required interval. One of the latter subjects was on a binge, which later resulted in his death due to medical complications. The other subject had moved and later was located out of the area.

Table 1 shows that the four SCT measures were independent, with two notable exceptions. At all three test administrations, noncompliance was related positively to specification of new behavior. At exit, duration of response was positively related to specification of new behavior, indicating that longer responses were likely to specify concrete behavior for most subjects only at the end of the inpatient phase of the treatment program.

Treatment effects on the SCT. There were no significant pretreatment group differences on the four dependent measures derived from the SCT. Treatment effects were analyzed by using the prescores to compute residuals of the exit and 3-month posttest scores, which then were subjected to a repeated measures analysis of variance (Huck & McLean, 1975). As Table 2 shows, for duration of response, the group main effect was significant. Planned comparisons indicate that at immediate retest, the skill-training group had a significantly longer duration of response than the discussion group, $t(34) = 1.69$, $p < .05$, and the control group, $t(34) = 2.44$, $p < .01$. As Figure 1

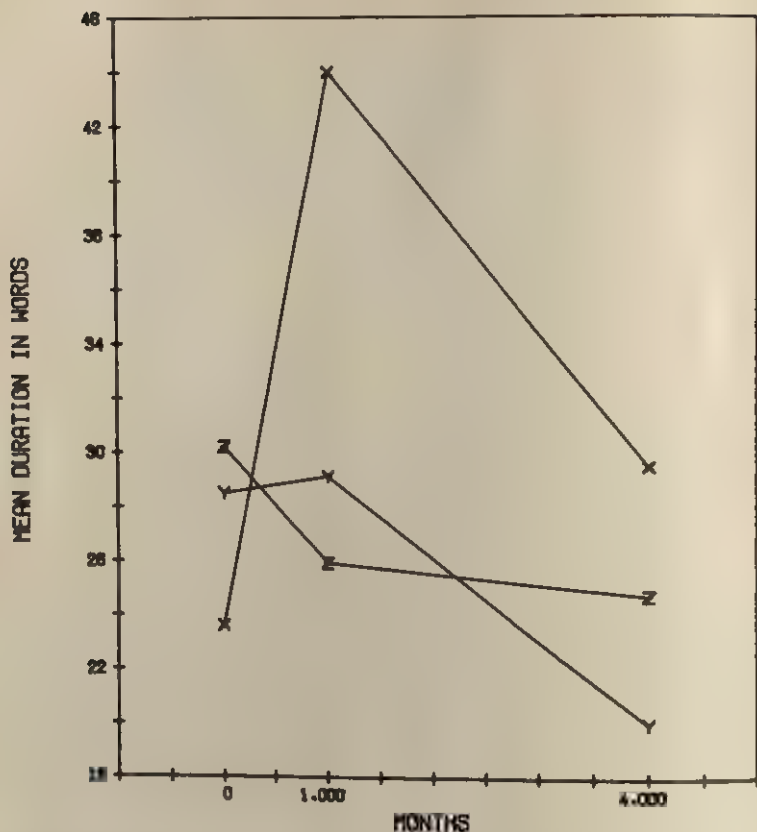


Figure 1. Response duration measured by the Situational Competency Test pre, post, and 3 months after training. (X = skill training; Y = discussion; Z = no treatment control.)

shows, by the 3-month posttest, these differences had diminished.

Specification of new behavior also showed a training effect, with the group main effect significant. Planned comparisons show that the skill-training group performed better on exit testing than both the discussion group, $t(34) = 2.44$, $p < .01$, and the control group, $t(34) = 1.69$, $p < .05$. By the 3-month posttest (see Figure 2), the skill-training group was still significantly different from the discussion group only, $t(34) = 1.69$, $p < .05$. No significant differences due to training were found on the latency or noncompliance measures.

Drinking Behavior and Social Adjustment

Posttreatment drinking behavior. Drinking behavior was assessed primarily by drinking disposition measures (Sobell & Sobell, 1973).

Days during the follow-up period were categorized as (a) voluntary abstinence; (b) forced abstinence (i.e., jailed or hospitalized); (c) controlled drinking, defined as any days during which 6 or fewer drinks were consumed or any isolated 1- or 2-day sequence when between 7 and 9 drinks were consumed; and (d) drunk days, defined as isolated days during which 10 or more drinks were consumed or as any day that was part of a period of more than 2 days during which 7-9 drinks were consumed. Total amount drunk and average drinking period length for those subjects who drank were computed for the year. Results exclude the skill-training subject who died.

Of the drinking disposition measures, too few subjects (nine in all) reported controlled drinking days for this category to be a useful outcome indicator. Preliminary analysis of the other drinking-related measures indicated that

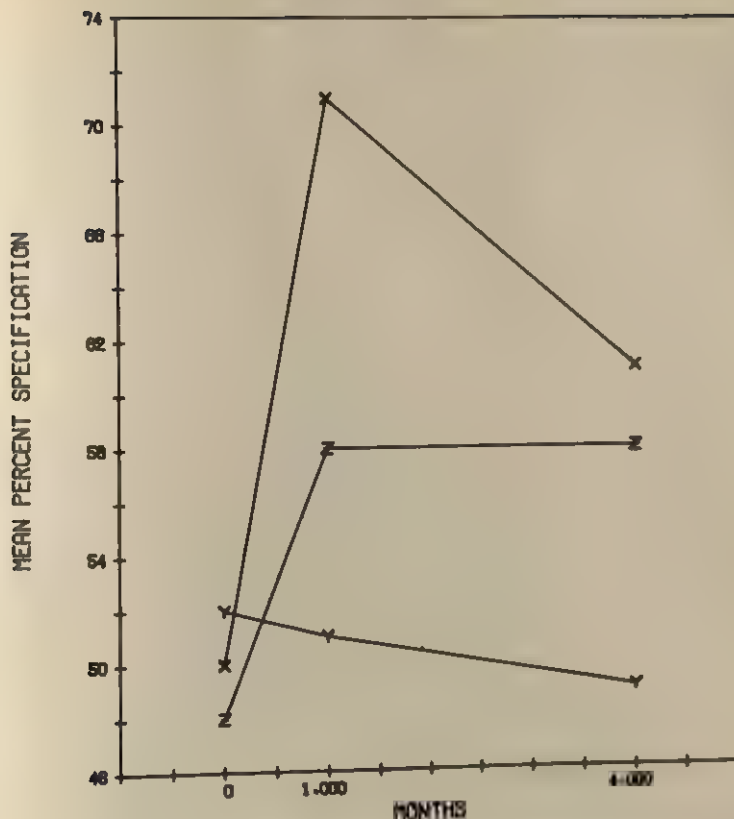


Figure 2. Specification of behavior measured by the Situational Competency Test pre, post, and 3 months after training. (X = skill training; Y = discussion; Z = no treatment control.)

within-cell variances were heterogeneous. The Bartlett-Box F (Winer, 1971) was significant for all measures except days voluntarily abstinent. Therefore, statistical analyses were performed on logarithmically transformed scores (Winer, 1971, p. 400).

First, the two control groups were compared on the drinking disposition measures using t tests. No significant differences were found, just as no differences between the two groups had been observed on the SCT. Thus these groups were pooled for comparison with the skill-training group. Since multiple dependent variables were involved, a multivariate analogue of the t test, Hotelling's T^2 (Winer, 1971), was applied to the following group of measures: days abstinent, days hospitalized, drunk days, total amount drunk, and average drinking period length; $T^2(5, 33) = 2.73$, $p < .05$. Having found the vectors of group means to be significantly different, individual t tests were

used to determine which measures accounted for the difference.

The skill-training and control groups were found to be significantly different on three of the five measures: days drunk, $t(37) = -2.21$, $p < .05$; total number of drinks, $t(37) = -2.01$, $p < .05$, and average drinking period length, $t(37) = -2.32$, $p < .05$. Table 3 gives untransformed means and standard deviations for these and the other outcome measures for the 1-year follow-up period. As the table shows, for the follow-up year, the skill-training group had an average number of days drunk one-sixth that of the pooled control group, drank one-fourth as much, and had an average drinking period length less than one-eighth as long.

To determine whether the improvement in drinking behavior generalized to areas of functioning that had not been specifically targeted by the intervention, two other self-

Table 3

Posttreatment Adjustment: 12-Month Outcome

Measure	Group ^a	<i>M</i>	<i>SD</i>	<i>df</i>	<i>t</i>
Days of controlled drinking	1	4.9	17.8		
	2	1.2	2.6		
Days drunk	1	11.1	14.3	37	-2.21*
	2	64.0	88.3		
Total no. drinks	1	399.8	507.8	37	-2.01*
	2	1,592.8	2,218.4		
Average drinking period length	1	5.1	6.9	37	-2.32*
	2	44.0	62.2		
Days hospitalized	1	44.6	96.9	37	.74
	2	28.2	42.0		
Days abstinent	1	298.6	100.8	37	.94
	2	266.7	101.8		
Days employed ^b	1	204.0	132.0	32	.55
	2	178.7	124.5		
Weekly aftercare meetings	1	29.8	18.8	37	.93
	2	24.2	17.8		

^a Group 1 (*n* = 14) received skill training; Group 2 (*n* = 25) subjects were in the discussion or non-additional-treatment groups. One Group 1 subject died after 5 months of an alcohol-aggravated illness following a 2-month period of substained drinking.

^b Five subjects were retired or pensioned (2 in Group 1 and 3 in Group 2).

* *p* < .05.

report measures of outcome were examined: job performance (retired and fully pensioned subjects excepted) and continued engagement in the ATP outpatient treatment for which the subject had contracted. These measures did not show significant group differences, suggesting that the effects of the intervention are relatively specific to drinking behavior.

Relapse Situation Category Analysis

A relapse was defined as the initiation of a drinking period lasting 1 or more drunk days. Successive relapses for the same individual were considered to be distinct when the subject reported different setting events for drinking periods, which were separated by at least 2 abstinent or controlled drinking days. For the year, 25 subjects (10 in the skill-training group, 6 discussion, and 9 control) reported a total of 55 relapses (*M* = 2.2, *SD* = 1.29, range = 1-6). Number of relapses for the subject groups did not differ significantly.

Relapse situations were categorized independently by two clinical psychology graduate students into the four categories used throughout the study, with 81% agreement. Disagreements were resolved by consensus. The largest

number of relapses were of the negative emotional state type (43%). The interpersonal temptation category was represented by 17% of the relapses, whereas frustration and anger situations and intrapersonal temptation each comprised 15.5%. Of the relapses, 9% were not classifiable either because of lack of sufficient information or failure to fit any of the four category descriptions.

Prediction of Outcome

To evaluate the ability of the SCT to predict posttreatment drinking behavior and social adjustment, a multiple regression approach incorporating drinking history and demographic measures was used. Due to the small number of subjects, the number of potential predictors was first reduced by correlating pretreatment variables with the 1-year outcome indices. Controlled drinking days were excluded since so few were reported. The three measures most highly related to each outcome index were entered into a stepwise regression analysis along with the three SCT exit measures: latency, duration, and specification of behavior. Noncompliance was not included because of its high collinearity with the specification measure.

Table 4
Prediction of 1-Year Outcome Variables

Outcome variable	Predictor	<i>r</i>	<i>F</i>	<i>R</i>	<i>R</i> ²	Overall <i>F</i>
Days employed	Pretreatment ^a	.35	4.42*	.35	.12	4.87*
	Post SCT latency	-.32	4.79*	.50	.25	
Day hospitalized	Post SCT latency	.50	11.81***	.50	.25	13.10****
	Drinking pattern ^b	.30	14.38****	.68	.47	
Weeks in aftercare	Previous treatments	.23	5.11*	.73	.54	7.01***
	Post SCT latency	-.39	6.62**	.39	.16	
Days abstinent	Marital status ^c	.38	6.41*	.53	.29	24.49****
	Post SCT latency	-.73	40.25****	.73	.53	
Days drunk	Marital status ^c	.27	4.65*	.76	.58	14.11****
	Post SCT latency	.51	12.74****	.51	.26	
Total amount drunk	Problem duration ^d	.34	11.70***	.67	.45	14.45****
	Post SCT latency	.52	13.25****	.52	.27	
Average drinking period length	Problem duration ^d	.34	11.71***	.67	.45	5.01**
	Drinks consumed ^e	.37	5.65*	.37	.14	
	Problem duration ^d	.33	3.96*	.47	.22	
	Post SCT latency	.35	4.08*	.55	.31	

Note. SSPS Stepwise Multiple Regression Program (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).
 SCT = Situational Competency Test.

^a Months employed on current job.

^b Dichotomous variable: Periodic/mixed or steady.

^c Dichotomous variable: Married or not.

^d Time elapsed since first alcoholism treatment.

^e During the 6 months prior to treatment.

* $p < .05$.

** $p < .01$.

*** $p < .005$.

**** $p < .001$.

Table 4 summarizes the significant findings giving simple and multiple correlation coefficients and *F* values for individual and combined predictors. For all outcome indices, the predictive ability of the SCT latency measure was comparable or superior to that of the most highly related demographic and drinking history measures. For instance, for days abstinent, latency of response to problematic situations accounted for 53% of the variance. For days employed and average drinking period length for which pretreatment measures entered the regression equations first, the latency measure still contributed significantly. This analysis establishes a relationship between ability to respond to problematic drinking-related situations on a verbal role-playing instrument and actual drinking-related behavioral functioning following treatment.

Discussion

The skill-training method successfully evaluated here views the client as an active organism

who can learn to cope with future problems. Approaches that *only* use verbal persuasion or substitute medication or supportive social groups for the treated addiction do not assist the client in reevaluating expectations of personal efficacy. Recent reviews (Bandura, 1977) have suggested that a person's expectations of success or failure in coping with situations are important determinants of behavior. Efficacy expectations are modifiable through corrective experience and training.

This study indicates that problem drinkers' responses to situations that present a high risk of relapse can be improved through training. Since the focus was on training problem-solving techniques, rather than a repertoire of specific responses, the testing situations were not the same as the situations used in training. The skill-training intervention produced longer and more specifically appropriate verbal behavior in response to these novel problem situations. These effects would not have been found unless generalization took place.

The measure of noncompliance demonstrated a ceiling effect. Both before and after treatment, subjects usually indicated that they would exercise control over problematic situations. The noncompliance of problem drinkers' verbal behavior in situations potentially leading to drinking probably reflects social desirability and demand characteristics of the assessment procedures rather than behavior *in vivo*. The sensitivity of the SCT to demand characteristics might have been accentuated by using the instructional set "what would you do" rather than "what are you going to do." On the other hand, pretesting indicated that such instructions were necessary to minimize frequent denial: "That will never happen to me!"

The apparent decline of training effects on the 3-month posttest may reflect in part that alcoholics typically appear for treatment after a prolonged period of heavy drinking. Consequently, they demonstrate cognitive impairment as evidenced by their low average Shipley CQ. The overall mean of 82 ($SD = 15.8$) is in the "moderately suspicious" range (Shipley, 1940). Research on the same population using Halstead-Reitan measures of cognitive functioning confirms this finding (Schau & O'Leary, 1977). The permanence of alcoholics' cognitive impairment and its relation to their drinking behavior is unknown (Rankin, 1975). Therapies incorporating concept learning (such as skill training) when used with recently sober problem drinkers whose cerebral functioning is at least temporarily impaired may require booster sessions administered on an outpatient basis for maximum enduring effect. The continuing beneficial impact on drinking outcome that was found may be due to the environmental support from significant others occasioned by successful coping with problematic situations encountered early in the follow-up period.

Evaluation of the finding that skill training had a beneficial effect over and above that of the regular treatment program on post-hospitalization adjustment is dependent on the validity of the self-report outcome measures. This study corroborated self-report with data systematically collected from collaterals. Comparison of subject and collateral reports suggests that certain measures of

drinking behavior are more reliable and less likely to elicit socially desirable answers than others. For instance, if a person goes on a binge and blacks out, the amount drunk during the binge will not be remembered accurately. In this respect, drinking disposition is an improvement over other measures.

Follow-up data indicated that during the posttreatment year the drinking behavior of the skill-training group and the control groups significantly diverged as measured by (a) days drunk, (b) amount drunk, and (c) length of drinking periods. Differences were found even though the regular treatment program was intensive. In the most successful behavioral study to date (Sobell & Sobell, 1973), much of the differential improvement in the functioning of experimental subjects as opposed to controls seemed to be accounted for by different amounts of controlled drinking (Lloyd & Salzberg, 1975). The fact that very few controlled drinking days were reported in the present study may be related to the treatment goal of abstinence. The skill-training sessions were conducted in accordance with the abstinence-oriented philosophy of the treatment program. The focus was on how not to drink rather than how to moderate drinking. However, these and other results (Armor, Polich, & Stambul, 1976) suggest that a majority of problem drinkers will resume drinking to some extent following treatment, no matter what that treatment consists of. It seems sensible not to encourage drinking but to stress prevention by preparing patients to cope with drinking situations when they inevitably arise.

A finding that supports the clinical utility of the SCT and the social competence approach to problem drinking is that the men in this sample who had shorter response latencies to role-played problematic drinking problem situations drank less, were employed more, and had more regular aftercare attendance following treatment. Of all variables, this indicator of problem-solving skill was most highly related to outcome. Avoiding problem drinking completely may be even more dependent on the ability to quickly generate an alternate response to drinking than on the precise content of the response, as modified by the skill-training procedure used here.

(Schwartz & Gottman, 1976, make a similar point regarding college students' responses to problematic situations.) Although latency is a difficult response characteristic to modify (Eisler, Hersen, & Miller, 1973; Hersen, Eisler, & Miller, 1974), future skill training might do well to incorporate explicit procedures to test whether this relationship is of a causal nature.

In conclusion we wish to stress three points.

(a) Skill training (even including booster sessions) *alone*, although a potentially effective component of multimodal treatment, probably would not be sufficient treatment for a rather socioeconomically and cognitively impaired population such as that studied here. (b) The treatment strategy evaluated here included not only the training of problem-solving skills but also behavioral rehearsal of specific responses (the verification phase). Determining which of the techniques used was most responsible for change requires further investigation. Future studies might draw on Bandura's (1977) self-efficacy conceptualization of behavior change when assessing individual elements of skill training and incorporating additional treatment components. (c) Further research on the characteristics and prevention of relapse in a variety of problem areas such as drug addiction, smoking, obesity, and crime is necessary and warranted. This study, in combination with Marlatt and Gordon's recent results (in press), suggests that different client populations find different types of relapse situations most problematic. A reliable taxonomy of relapse situations related to the characteristics of different client populations, together with further refinement of skill-training procedures, would help in the attempt to match the treatment to the person.

Reference Note

1. Hollingshead, A. B. *Two-factor index of social position*. Unpublished manuscript, Yale University, 1957.

References

- Allman, L. R., Taylor, H. A., & Nathan, P. E. Group drinking during stress: Effects on drinking behavior, affect, and psychopathology. *American Journal of Psychiatry*, 1972, 129, 669-678.
- Armor, D. J., Polich, J. M., & Stambul, H. B. *Alcoholism and treatment* (R-1739-NIAAA). Santa Monica, Calif.: The Rand Corporation, 1976.
- Arnkoff, D. B., & Stewart, J. The effectiveness of modeling and videotape feedback on personal problem solving. *Behaviour Research and Therapy*, 1975, 13, 127-133.
- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 1977, 84, 191-215.
- Curran, J. P. Skills training as an approach to the treatment of heterosexual-social anxiety: A review. *Psychological Bulletin*, 1977, 84, 140-157.
- D'Zurilla, T. J., & Goldfried, M. R. Problem solving and behavior modification. *Journal of Abnormal Psychology*, 1971, 78, 107-126.
- Eisler, R. M., Hersen, M., & Miller, P. M. Effects of modeling on components of assertive behavior. *Journal of Behavior Therapy and Experimental Psychiatry*, 1973, 4, 1-6.
- Eisler, R. M., Hersen, M., Miller, P. M., & Blanchard, E. B. Situational determinants of assertive behaviors. *Journal of Consulting and Clinical Psychology*, 1975, 43, 330-340.
- Eisler, R. M., Miller, P. M., & Hersen, M. Components of assertive behavior. *Journal of Clinical Psychology*, 1973, 29, 295-299.
- Finch, B. E., & Wallace, C. J. Successful interpersonal skills training with schizophrenic inpatients. *Journal of Consulting and Clinical Psychology*, 1977, 45, 885-890.
- Foy, D. W., Miller, P. M., Eisler, R. M., & O'Toole, D. H. Social-skills training to teach alcoholics to refuse drinks effectively. *Journal of Studies on Alcohol*, 1976, 37, 1340-1345.
- Goldsmith, J. B., & McFall, R. M. Development and evaluation of an interpersonal skill-training program for psychiatric inpatients. *Journal of Abnormal Psychology*, 1975, 84, 51-58.
- Hamburg, S. Behavior therapy in alcoholism: A critical review of broad-spectrum approaches. *Journal of Studies on Alcohol*, 1975, 36, 69-87.
- Hersen, M., Eisler, R. M., & Miller, P. M. An experimental analysis of generalization in assertive training. *Behaviour Research and Therapy*, 1974, 12, 295-310.
- Higgins, R. L., & Marlatt, G. A. Fear of interpersonal evaluation as a determinant of alcohol consumption in male social drinkers. *Journal of Abnormal Psychology*, 1975, 84, 644-651.
- Huck, S. W., & McLean, R. A. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, 1975, 82, 511-518.
- Lawrence, P. S. The assessment and modification of assertive behavior (Doctoral dissertation, Arizona State University, 1970). *Dissertation Abstracts International*, 1970, 31, 396B. (University Microfilms No. 70-11, 888).
- Lloyd, R. W., & Salzberg, H. C. Controlled social drinking: An alternative to abstinence as a treatment

- goal for some alcohol abusers. *Psychological Bulletin*, 1975, 82, 815-842.
- Marlatt, G. A. The Drinking Profile: A questionnaire for the behavioral assessment of alcoholism. In E. J. Mash & L. G. Terdal (Eds.), *Behavior therapy assessment: Diagnosis and evaluation*. New York: Springer, 1976.
- Marlatt, G. A. Craving for alcohol, loss of control, and relapse: A cognitive behavioral analysis. In P. E. Nathan, G. A. Marlatt, & T. Löberg (Eds.), *Alcoholism: New directions in behavioral research and treatment*. New York: Plenum Press, 1978.
- Marlatt, G. A., & Gordon, J. R. Determinants of relapse: Implications for the maintenance of behavior change. In P. Davidson (Ed.), *Behavioral medicine, changing health lifestyles*. New York: Brunner-Mazel, in press.
- McFall, R. M., & Lillesand, D. B. Behavior rehearsal with modeling and coaching in assertion training. *Journal of Abnormal Psychology*, 1971, 77, 313-323.
- McFall, R. M., & Marston, A. R. An experimental investigation of behavioral rehearsal in assertive training. *Journal of Abnormal Psychology*, 1970, 76, 295-303.
- McFall, R. M., & Twentyman, C. T. Four experiments on the relative contributions of rehearsal, modeling and coaching to assertion training. *Journal of Abnormal Psychology*, 1973, 81, 199-218.
- Miller, P. M., Hersen, M., Eisler, R. M., & Hilsman, G. Effects of social stress on operant drinking of alcoholics and social drinkers. *Behaviour Research and Therapy*, 1974, 12, 67-72.
- Nathan, P. E. Alcoholism. In H. Leitenberg (Ed.), *Handbook of behavior modification and behavior therapy*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. *SPSS: Statistical package for the social sciences*. New York: McGraw-Hill, 1975.
- Rankin, J. G. (Ed.). *Alcohol, drugs and brain damage*. Toronto, Canada: Addiction Research Foundation, 1975.
- Rich, A. R., & Schroeder, H. E. Research issues in assertiveness training. *Psychological Bulletin*, 1976, 83, 1081-1096.
- Schau, E. J., & O'Leary, M. R. Adaptive abilities of hospitalized alcoholics and matched controls: The brain-age quotient. *Journal of Studies on Alcohol*, 1977, 38, 403-409.
- Schwartz, R. M., & Gottman, J. M. Toward a task analysis of assertive behavior. *Journal of Consulting and Clinical Psychology*, 1976, 44, 910-920.
- Shipley, W. C. *Shipley-Hartford Retreat Scale: Manual of directions and scoring key*. Hartford, Conn.: Hartford Retreat, 1940.
- Sobell, M. B., & Sobell, L. C. Individualized behavior therapy for alcoholics. *Behavior Therapy*, 1973, 4, 49-72.
- Toomey, M. Conflict theory approach to decision making applied to alcoholics. *Journal of Personality and Social Psychology*, 1972, 24, 199-206.
- Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Received September 15, 1977 ■

Evaluating Alcoholism Treatment Programs: An Integrated Approach

Ruth C. Cronkite and Rudolf H. Moos

Department of Psychiatry and Behavioral Sciences
Social Ecology Laboratory, Stanford University School of Medicine
and Palo Alto Veterans Administration Hospital, Palo Alto, California

This article examines the interrelationships among five major sets of variables (social background, intake symptoms, program type, treatment experiences, and perceptions of the environment) that are related to posttreatment functioning of alcoholic patients (alcohol consumption, rating of drinking problem, physical impairment, and occupational functioning). The sample consisted of 429 patients selected from five different treatment programs. The relative importance of each set of variables as predictors of outcome was estimated by constructing block variables, using path analyses, and partitioning the explained variance. The results showed that (a) the combined explanatory power of the program-related variables is considerably more than would be expected from previous research; (b) the importance of patient background relative to intake symptoms varies with the outcome criterion being used; (c) both the treatment experiences and the patient's perceptions of the treatment environment are strong predictors of outcome; and (d) a substantial proportion of the explained variance is shared between patient-related and program-related variables, suggesting important patient-program selection and congruence effects.

One of the major issues in longitudinal studies of alcoholic patients is assessing the relative importance of patient background and treatment programs in determining outcome. Although contradictory findings have been reported, previous research has generally suggested that patient characteristics at intake are most strongly related to outcome and that treatment programs have little effect once sociodemographic and functioning characteristics at intake are taken into account (Armor, Polich, & Stambul, 1976;

Craft, Sheehan, Driggers, & DuBois, 1975; Gerard & Saenger, 1966; Pokorny, Miller, & Cleveland, 1968; Ruggels, Armor, Polich, Mothershead, & Stephen, 1975).

A related issue involves the relative contributions of different types of patient characteristics—in particular, the extent to which outcome is related to social background on the one hand and drinking symptoms at intake on the other. No clear-cut pattern of findings has been reported on the relative importance of social background variables (such as socioeconomic or marital status) compared to intake symptoms (such as alcohol consumption or behavioral or psychological impairment at intake) in predicting outcome (Armor et al., 1976; Craft et al., 1975; Ruggels et al., 1975).

Inferences pertaining to both of these issues have been primarily based on examination of the increments in explained variance (Armor et al., 1976; Bromet, Moos, Bliss, & Wuthmann, 1977; Craft et al., 1975; Ruggels et al., 1975). This method is asymmetric

The research was supported in part by National Institute for Alcohol Abuse and Alcoholism Grant AA02863 and by Veterans Administration Research Funds. The article is based on a presentation to the Evaluation Research Session, Western Social Science Association meetings, Denver, Colorado, April 1977. Ruth Cronkite is also affiliated with Mills College, Oakland, California.

Requests for reprints should be sent to Rudolf Moos, Social Ecology Laboratory, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California 94305.

in that it attributes variance shared by two or more sets of variables to the set that is entered first in the regression analysis (such as patient background variables), and thus may overestimate the contribution of the first set by crediting it with both its unique and shared variance. In contrast, the increment in explained variance attributed to the set of variables added last (such as program-related variables) represents only the explained variance that is unique to that set of variables, not the variance that is shared with variables that have been entered earlier. Consequently, previous inferences about the relative effects of program-related variables, social background characteristics, and intake symptoms on outcome may be misleading.¹ Such inferences have important policy implications for alcoholism treatment. For example, if the unique variance attributed to program-related variables is small, and the variance that is shared with patient background variables is attributed only to background characteristics, then researchers may conclude that treatment effects are negligible and thus recommend less expensive and more uniform treatment programs.

Another unresolved question raised by longitudinal studies of alcoholics is the role of different types of program-related variables in predicting outcome. The relationship of treatment variations to outcome has been approached in several ways. Armor et al. (1976) and Kissin, Platz, and Su (1970) focused on the type and amount of treatment both within and across programs. Bromet et al. (1977) examined the effect of level of participation on outcome in several different treatment programs. In addition to studying the effects of a variety of treatment experiences and treatment programs, Bromet, Moos, and Bliss (1976) have focused on another dimension of program-related variables, a patient's perceptions of the treatment environment. This approach is based on research which suggests that the social environments of psychiatric and correctional programs may be important factors in influencing outcome (Ellsworth, Maroney, Klett, Gordon, & Gunn, 1971; Moos, 1974b). Each of these aspects of treatment variations has been studied separately, but their effects relative

to each other have not been examined within a single study.

The purpose of this article is to develop an integrated approach to studying alcoholism treatment programs that will facilitate clarification of the following issues: (a) What are the interrelationships among patient social background variables, intake symptoms, treatment programs, treatment experiences, perceptions of the treatment environment, and outcome? (b) What is the relative importance of program-related variables compared to patient background characteristics in predicting outcome? (c) What is the relative importance of a patient's social background compared to intake symptoms in predicting outcome? (d) What are the roles of different aspects of treatment variations in predicting outcome?

A Model of Treatment Outcome

In most longitudinal studies of alcoholic patients, the dependent variables are one or more outcome criteria related to posttreatment functioning, such as rehospitalization, alcohol consumption, and occupational, physical, and psychosocial functioning. The independent variables vary across studies, depending on their focus. From reviewing previous research, the independent variables can be divided into "blocks," labeled *Blocks 1, 2, 3, 4, and 5*. Block 1 consists of a set of sociodemographic variables known to be related to drinking patterns, such as age, sex, ethnicity, marital status, and socioeconomic status. Block 2 includes a patient's drinking symptoms at intake, more specifically, the type and severity of alcoholism-related characteristics, such as alcohol consumption, drinking patterns, physical impairment, and behavioral impairment. Blocks 3, 4, and 5 refer to program-related variables. Block 3 includes the type of program; Block 4 includes the amount of various treatment experiences, such as therapy sessions, Alcoholics Anonymous (AA) meetings, antabuse, and

¹ See Newton and Spurrell (1967a, 1967b) and Mood (1971) for a complete technical discussion of partitioning the sum of squares in regression analysis and Coleman (1975) for a discussion of this issue when applied to school effects.

so forth, in which the patient participates; and Block 5 includes the patient's perceptions of the program environment.

As mentioned earlier, researchers have focused on comparing Blocks 1 and 2 with Blocks 3 and 4 (and to a lesser extent Block 5) in an attempt to assess the effect on outcome of program-related variables compared to patient characteristics at intake (Armor et al., 1976; Bromet et al., 1977; Ruggels et al., 1975). Some of these investigators have also compared Block 1 variables with Block 2 variables to focus on the relative contributions of social background and intake symptoms (Armor et al., 1976; Ruggels et al., 1975) in explaining outcome.

The use of block variables not only serves to group conceptually similar variables, but it also allows one to formulate a model that summarizes the hypothesized causal interrelationships among the block variables, as shown in Figure 1. Blocks 1 and 2 represent characteristics of a patient at the time of entering a program. Although some of the Block 1 variables can be regarded as antecedents of the onset of alcoholism (Armor et al., 1976), they also reflect a patient's social characteristics at the time of intake to treatment and can thus be specified as correlated with intake symptoms (represented by the bidirectional line, r_{12}). Not only are both social background (Block 1) and intake symptoms (Block 2) related to a patient's outcome after treatment, as shown by the paths p_{61} and p_{62} , but they are also important determinants of the type of program that a patient enters (e.g., patients who enter private treatment programs tend to be older, married, and to have higher income, educational, and occupational levels; see Bromet et al., 1977), as specified by p_{32} and p_{31} .

The treatment experiences that a patient receives are almost entirely determined by the program that a patient enters, which is specified by the path p_{43} . Bromet et al. (1976) have shown that patient characteristics such as social background and intake symptoms are unrelated to the treatment experiences received by patients. Consequently, the paths from Blocks 1 and 2 to Block 4 are not in-

cluded in the model (i.e., p_{41} and p_{42} were constrained to be zero).²

Similarly, since the environments of treatment programs vary considerably in the degree of support, autonomy, clarity, staff control, and so on, the program is expected to be an important determinant of a patient's perceptions of the treatment environment, represented by p_{53} (Bromet et al., 1976). In addition, it is possible that patients' perceptions of the treatment program may be affected by their characteristics at intake, as specified by the path p_{51} . All of the program-related variables, Blocks 3, 4, and 5, are expected to directly affect outcome, as indicated by Paths p_{63} , p_{64} , and p_{65} .

This model summarizes the hypothesized causal interrelationships among the sets of variables related to outcome. It also displays the causal reasoning behind the practice of entering blocks of variables sequentially into the regression analyses in which social background (Block 1) and intake symptoms (Block 2) precede the type of treatment program, and, in turn, the treatment program (Block 3) precedes the other program-related variables, treatment experiences (Block 4), and perceptions of the environment (Block 5). Furthermore, this model allows one to establish both the direct and indirect paths through which each block of variables can affect and be affected by other blocks of variables. For example, social background not only has a direct effect on outcome that is independent of the other variables in the model, via p_{61} , but it can also indirectly affect outcome via intervening variables, such as the program ($p_{63}p_{31}$). That is, the compound path $p_{63}p_{31}$ represents the indirect effect of social background that is mediated by the program (or the effect of social background that is shared with the program effects). Other indirect paths include the compound paths in the model that link two variables via one or more intervening variables (e.g., $p_{65}p_{51}$, $p_{65}p_{53}p_{31}$, and $p_{64}p_{43}$). When all of the direct

² Analyses that were undertaken with a fully recursive model revealed that the omitted paths in Figure 1 (p_{41} , p_{42} , p_{52} , and p_{54}) were all close to zero, thus providing empirical support for the more constrained model.

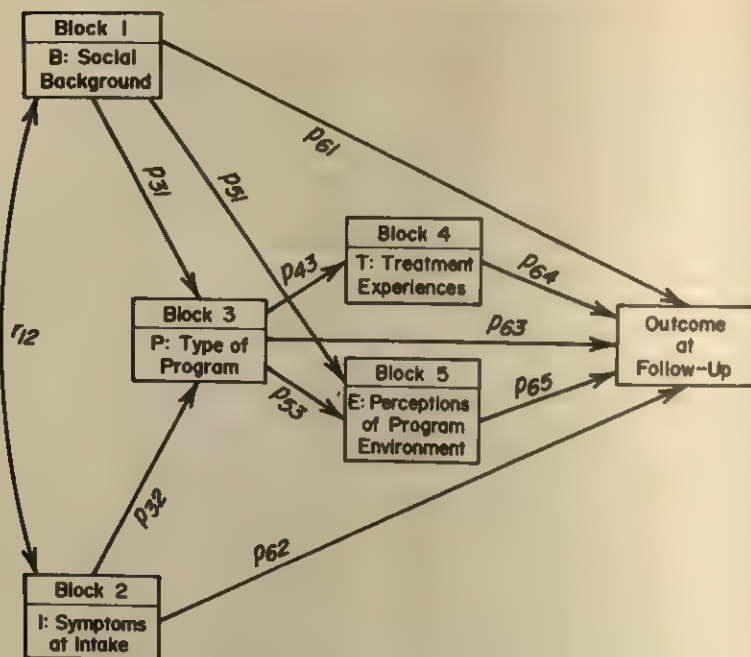


Figure 1. A model of treatment outcome.

and indirect effects are taken together with information on the unique and joint explained variance, it is possible to assess more accurately the extent to which each set of block variables contributes to explaining treatment outcome.

Method

Sample, Programs, and Data

The sample consisted of 429 patients from five residential alcoholism programs. These facilities, which were selected because they are representative of the major types of treatment settings for alcoholics, include (a) a Salvation Army program offering milieu therapy and vocational rehabilitation to skid-row alcoholics; (b) a public hospital-based facility offering milieu and group treatment and anti-anxiety medicines and sedatives to low-income patients; (c) a county-funded halfway house operating as a therapeutic community with individual and group therapy and recreational activities; (d) a private aversion-conditioning medication-oriented program treating middle-class and upper-middle-class patients; and (e) a private milieu-oriented program emphasizing group and family therapy (as well as medication) for middle-class and upper-middle-class patients. (For more detailed information, see Bromet et al., 1976.)

Patients in the study included approximately 80%

of all admissions during a 10-month average period. Most of the nonparticipants were those who dropped out of the program within a few days. Unlike many other studies, the sample is representative of patients from a wide range of socioeconomic characteristics such as education, age, income, employment status, and marital status.

The data consisted of (a) a background information form administered to patients shortly after admission to a program; (b) a treatment experiences form completed by a staff member after a patient left the program; (c) the Community-Oriented Programs Environment Scale (COPES) administered to patients about 2 weeks after admission to measure their perceptions of the program environment; and (d) a follow-up information form completed by the patient approximately 6 months after discharge.³

Background information form. The following data on a patient's social background and drinking symptoms at intake came from the background information form:

1. Social background (five items): age, sex, marital status (married, not married), ethnic group (white, nonwhite), and highest grade of school completed.
2. Drinking symptoms at intake: (a) alcohol consumption, the quantity of alcohol in ounces of ethanol from beer, wine, and hard liquor consumed on a typical drinking day (Armor et al., 1976); (b) physi-

³ Although the validity of data based on self-report of alcoholics is controversial, Sobell and Sobell (1975) have shown that it may be relatively reliable and valid.

cal impairment, a subscale derived from the mean of 10 items rated on 5-point scales (from never to often), referring to how often patients experience delirium tremens or shakes, memory lapses or black-outs, dry heaves or cold sweat, upset stomach, dizzy spells, and so forth; (c) subjective rating of drinking problem, ranging from 1 for no problem to 5 for quite often a problem; and (d) occupational functioning, which refers to whether the patient had been unemployed during the last 6 months prior to admission (yes/no).

Treatment experiences. Together the programs offered a wide range of treatment experiences. Two programs (the hospital-based program and the private milieu-oriented program) offered ataractic medication, psychotherapy, and educational interventions. Two others (the Salvation Army program and the public hospital-based program) emphasized primarily psychotherapeutic and educational interventions, whereas the aversion-conditioning program relied almost entirely on its conditioning schedule (using emetic and ataractic drugs) and minimized other types of treatment experiences. All programs offered AA sessions, even though they varied in the extent to which they were emphasized. A total of 12 types of treatment experiences were offered across the five programs: antianxiety medications, sedatives, vitamins, antabuse, psychotherapy sessions, house meetings, AA sessions, educational lectures and films, recreational activities, other informal group activities, and (at the Salvation Army program only) Sunday worship and spot jobs. The amount of treatment, level of participation, and length of stay varied considerably among the patients within each program.

Perceptions of the environment. The COPEs was used to measure a patient's perceptions of the program environment. The COPEs is a scale designed to measure 10 dimensions of the treatment environment by having patients respond to 100 true-false items, which fall into 10 subscales. Three of the subscales measure personal relationship dimensions (Involvement, Support, and Spontaneity). Four subscales assess personal development or treatment program dimensions, such as the extent to which patients are encouraged to be self-sufficient and independent (Autonomy), to prepare for the future (Practical Orientation), and to have insight into their problems (Personal Problem Orientation and Anger and Aggression). The last three subscales (Order and Organization, Program Clarity, and Staff Control) assess system maintenance dimensions (see Moos, 1974a, 1974b).

Follow-up information form. The follow-up evaluation, conducted approximately 6-8 months after discharge, was obtained using a questionnaire identical in content to the background information form administered at intake. The follow-up form was completed by 429 patients (87% of the 494 patients available for follow-up). Four outcome criteria were selected to assess major dimensions of posttreatment functioning: alcohol consumption in ounces of ethanol, subjective rating of drinking problem, physical

impairment, and occupational functioning. These four dimensions correspond to the four intake symptoms drawn from the background information form.

Procedure

There were three major steps in carrying out the analyses: (a) construction of "block" variables; (b) the use of path analysis to compute the direct, indirect, and total effects among the block variables in the model; and (c) the partitioning of the sum of squares of the regression analysis into the unique variance attributed to each block variable and the joint variance shared by combinations of block variables.

Construction of block variables. Coleman (1975) has suggested a procedure for constructing a single composite or block variable from a set of variables that are similar conceptually (e.g., variables such as age, sex, marital status, education, and ethnicity can jointly define a composite variable called "social background"). A composite variable is computed as a weighted sum of the variables that compose a block. The weights are the regression coefficients obtained when the set of variables are the only independent variables used to predict the dependent variable. For example, a single variable representing social background would be the weighted sum of the various social background variables, the weights being the regression coefficients obtained when a particular outcome criterion (such as alcohol consumption) is the dependent variable. Such a composite variable no longer has a natural metric, but it still ranges from low to high levels of social background, where a high level refers to those combinations of social background characteristics that are associated with better outcome.

Four composite variables representing social background (Block 1), program type (Block 3), treatment experiences (Block 4), and perceptions of the program environment (Block 5) were constructed separately for each of the four outcome criteria. That is, by using a particular outcome criterion, such as alcohol consumption, regression coefficients were obtained separately for each set of variables that composed a block variable and were used as weights to construct that block variable. Four sets of four block variables were computed from regression coefficients obtained for each of the four outcome criteria.⁴

Instead of constructing a composite intake symptom variable, the intake symptom corresponding to

⁴To compare the regression results using the newly constructed block variables with a full regression model, the coefficients of the block variables were multiplied by each of the coefficients actually obtained from a full regression model. This comparison confirmed that the general pattern of results is similar even though the composite variable model is more restricted (in the way that it allows the variables within and across models to be cross-correlated).

the outcome criterion in each model was selected. Using a single intake symptom in this case serves to clarify the relationship between the same type of alcoholism-related characteristics at two points in time.⁵

Calculation of direct, indirect, and total effects. By using path analysis with recursive models, it is possible to calculate the direct, indirect, and total effects of each block variable (and the single intake symptom) on the other variables in the model. The general method involves estimating the structural equations that correspond to a recursive model, such as the one presented in Figure 1. Alwin and Hauser (1975) suggested a general method for decomposing total effects into their direct and indirect effects through the estimation of successive reduced-form equations.

The model illustrated in Figure 1 can be represented by the following set of equations:

$$X_3 = p_{31}X_1 + p_{32}X_2 + e, \quad (1)$$

$$X_4 = p_{43}X_3 + u, \quad (2)$$

$$X_5 = p_{51}X_1 + p_{53}X_3 + v, \quad (3)$$

$$X_6 = p_{61}X_1 + p_{62}X_2 + p_{63}X_3 + p_{64}X_4 + p_{65}X_5 + w, \quad (4)$$

where X_1 = the block variable representing social background, X_2 = the selected intake symptom, X_3 = the block variable representing "program type," X_4 = the block variable representing "treatment experiences," X_5 = the block variable representing "perceptions of the program environment," X_6 = the selected outcome criterion, and e , u , v , and w are random error terms.

Consistent with conventional notation in path analysis, direct effects are represented by p s (e.g., p_{31} is the direct effect of social background on outcome). Total effects are represented by q s (e.g., q_{31} is the total effect of social background on outcome). A total effect is the sum of the direct effect (e.g., p_{31}) and any indirect effects via intervening variables (e.g., $q_{31} = p_{31} + p_{32}p_{43} + p_{32}p_{43}p_{51} + p_{32}p_{43}p_{51}p_{61} + p_{32}p_{43}p_{51}p_{62}$). The total effect is equal to the direct effect if there are no indirect effects (e.g., $q_{31} = p_{31}$), and it is equal to an indirect effect if there is no direct path (e.g., $q_{41} = p_{43}p_{31}$).

All of the direct, indirect, and total effects corresponding to the model presented in Figure 1 and specified by Equations 1-4 were estimated for the four outcome criteria. This decomposition of the total effects allows for a detailed examination of the interrelationships among the variables related to outcome.

Path coefficients (standardized regression coefficients), rather than unstandardized regression coefficients, were used because the block variables do not have a natural metric. Unstandardized coefficients are difficult to interpret when it is unclear what is meant by a unit increase in a block variable. This problem is avoided with path coefficients, since they represent changes in standard deviation units.

Partitioning the sums of squares. As noted above, previous research on alcoholism treatment programs,

which compares the amount of explained variance accounted for by different types of variables, has been asymmetric because of the sequential order in which variables are entered into the regression analysis. That is, the variance associated with the variable entered first includes not only the explained variance that is unique to that variable but also the variance that it shares with other variables that are entered later in the analysis. Consequently, the increment in explained variance reflects only the unique variance associated with the added variable, since any variance that it shares with preceding variables has already been accounted for.

Newton and Spurrell (1967a, 1967b) and Mood (1971) have outlined a procedure for making more symmetric comparisons by calculating the unique and shared variance accounted for by each variable or combination of variables in a regression equation. This analysis was carried out for each of the four outcome criteria by varying the sequential order in which the four block variables and the single intake variable were included in the regression. The explained variance was then partitioned into the variance attributed uniquely to each block variable (or the single intake variable) and the variance shared by the various combinations of variables.⁶

Results

Three sets of analyses are presented: (a) the interrelationships among the patient background and program-related variables; (b) a comparison of total, direct, and indirect effects of patient background and program-related variables on outcome; and (c) an examination of the unique and shared explained variance attributed to the variables used to predict outcome.

The path diagrams of the four selected outcome criteria are shown in Figures 2-5. The path coefficients corresponding to the effect of the patient background variables (social background and the intake symptom) on the program-related variables (program type, treatment experiences, and perceptions of the treat-

⁵ Some analyses were conducted using composite intake symptom variables. The results were very similar to those that used only a single intake symptom, with a tendency for the composite variable to have a slightly stronger effect on outcome.

⁶ Since any analysis may be subject to particular biases and instabilities, both path analysis and partitioning of the explained variance are used. This allows one to have more confidence in the results and minimizes the likelihood that the results are due to possible instabilities or biases in the regression weights.

ment environment) are listed in Table 1, along with estimates of their total and indirect effects. The rest of the path coefficients shown in Figures 2-5, as well as the corresponding total and indirect effects, are displayed in Table 2. These estimates reflect the effect of the patient background variables (social background and intake symptom) and the program-related variables (program type, treatment experiences, and perceptions of the treatment environment) on outcome.

Relationship of Patient Characteristics to Program-Related Variables

Determinants of program selection. The estimates of p_{81} shown in Figures 2-5 (also listed in Table 1) for the four outcome criteria show that the composite background variable is an important determinant of the type of program that a patient enters. (All are statistically significant at the .05 level.) The path coefficients across three of the models are very close, ranging from .463 in the physical concomitants model to .559 in the alcohol consumption model. The slightly lower coefficient in the occupational functioning model may result from the nature of this particular outcome criterion [i.e., the effect of the intake characteristic, occupational functioning (p_{32}) probably has accounted for some of the effect of social background, because of its similarity to the other social background characteristics].

The estimates of the effects of the other intake functioning characteristics (p_{32}) indicate that these three alcoholism-related symptoms are not consistently strong determinants of the type of program that a patient enters, as shown by coefficients ranging from $-.150$ to $.015$. The two intake variables that had relatively strong effects (occupational functioning and physical concomitants) are those that were most highly correlated with social background.

Determinants of treatment experiences. The only variable hypothesized to have a direct effect on treatment experiences is the program type, and, in fact, the estimates of the effect of the program (p_{43}) are relatively large (and statistically significant), ranging from .460 to .689 across all four models (see Table

1 and Figures 2-5). This finding is consistent with expectations, since the programs almost entirely determine the type of treatment experiences that a patient has.

Although neither social background nor the intake symptom were hypothesized to have a direct effect on treatment experiences, their indirect effects via the program type were estimated and are listed in Table 1. The indirect effects of the intake symptoms via the program type ($p_{43}p_{32}$) are relatively small, as shown by estimates ranging from $-.089$ to $.007$. In contrast, social background seems to have a stronger indirect effect, with estimates of $p_{43}p_{31}$ of .124 and .360. This implies that patients with a "higher" level of social background tend to enter and/or more actively participate in programs offering those treatment experiences that are associated with better outcome.

Determinants of perceptions of the environment. The estimates of p_{51} in Table 1 show that social background has a positive effect on a patient's perceptions of the program environment (with the exception of the model of occupational functioning), suggesting that patients with higher levels of social background characteristics tend to perceive the environment slightly more positively. The decomposition of the total effect of social background on perceptions of the environment (q_{51}) indicates that most of the total effect is due to the indirect positive effect of social background via the program type ($p_{53}p_{31}$), implying that those patients with higher levels of social background characteristics enter programs that provide more positive treatment environments. Similar to the effects of the program type on treatment experiences, the direct effects of the program type on perceptions of the environment (p_{53}) are relatively close and large (ranging from .434 to .552), indicating that the major determinant of a patient's perceptions is the program that the patient is in.

Relationship of Patient Background and Program-Related Variables to Outcome

Table 2 displays the direct, indirect, and total effects of each of the five variables hypothesized to affect outcome. The results show

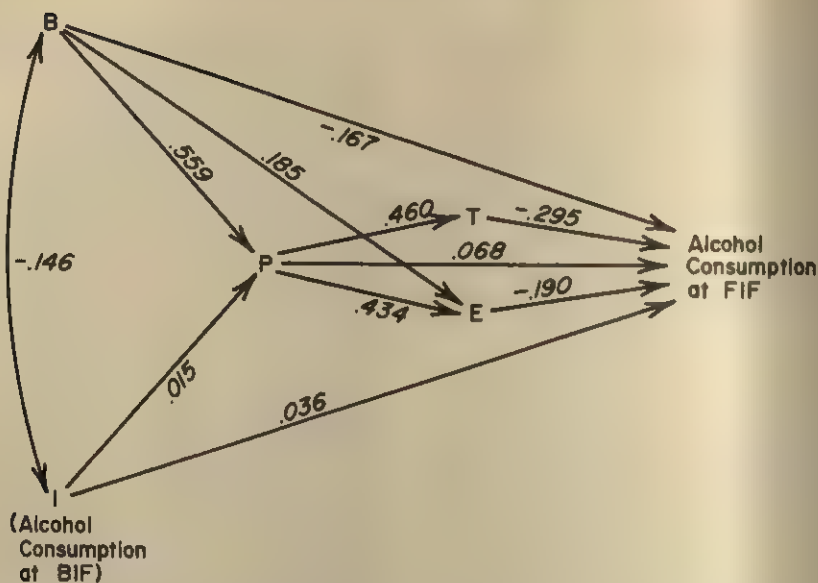


Figure 2. Estimation of path coefficients for alcohol consumption. (FIF = follow-up information form; BIF = background information form; B = background; I = intake; P = program; T = treatment experiences; E = treatment environment.)

some clear-cut consistencies across the models. The estimates of p_{61} (three of which are statistically significant) suggest that the higher the level of social background, the less severe the alcoholism-related symptoms are at fol-

low-up. Similarly, there is a relatively strong association between a patient's intake symptoms and the corresponding outcome criterion at follow-up, indicated by the estimates of p_{62} (except for a weak effect in the alcohol

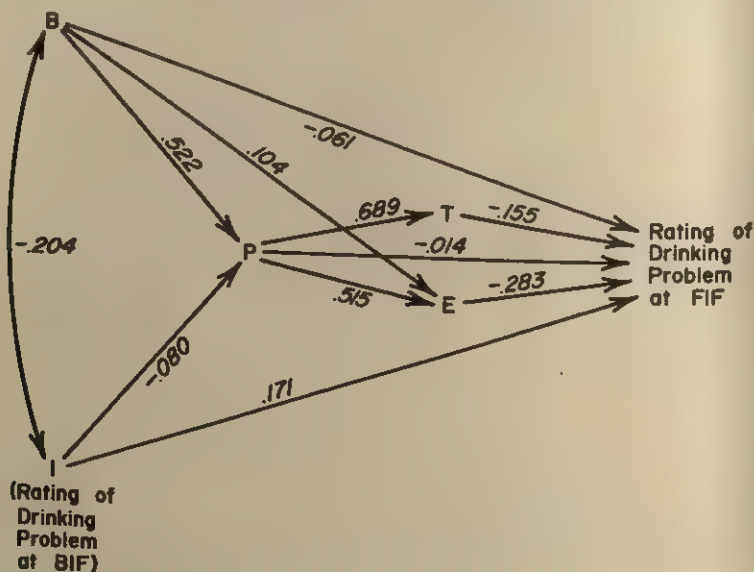


Figure 3. Estimation of path coefficients for rating of drinking problem. (FIF = follow-up information form; BIF = background information form; B = background; I = intake; P = program; T = treatment experiences; E = treatment environment.)

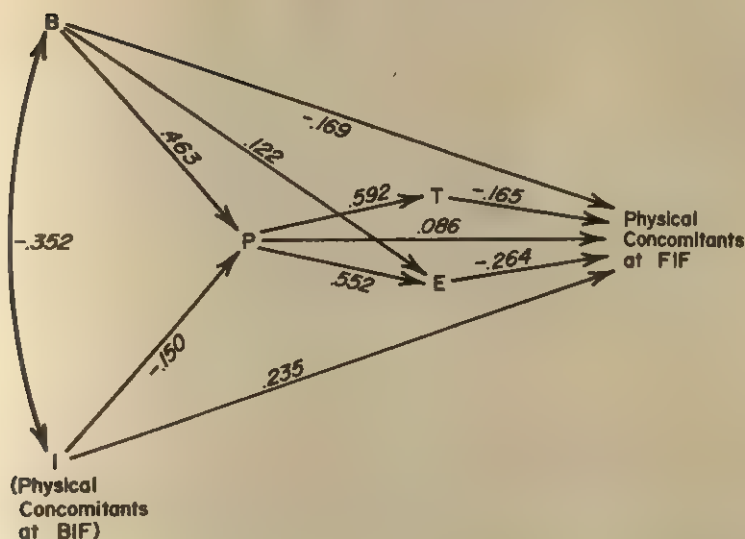


Figure 4. Estimation of path coefficients for physical concomitants. (FIF = follow-up information form; BIF = background information form; B = background; I = intake; P = program; T = treatment experiences; E = treatment environment.)

consumption model). In fact, for three of the outcome criteria, the intake symptoms had slightly stronger direct effects than social background. The indirect effects of social background that are mediated by the program-related variables together represent between 11% and 73% of their total effect,

whereas the indirect effects of the intake symptom together represent between 7% and 23% of their total effect. These results indicate that in three of the models, a substantial proportion of the total effect of social background is via indirect effects that are shared with the program-related variables. In con-

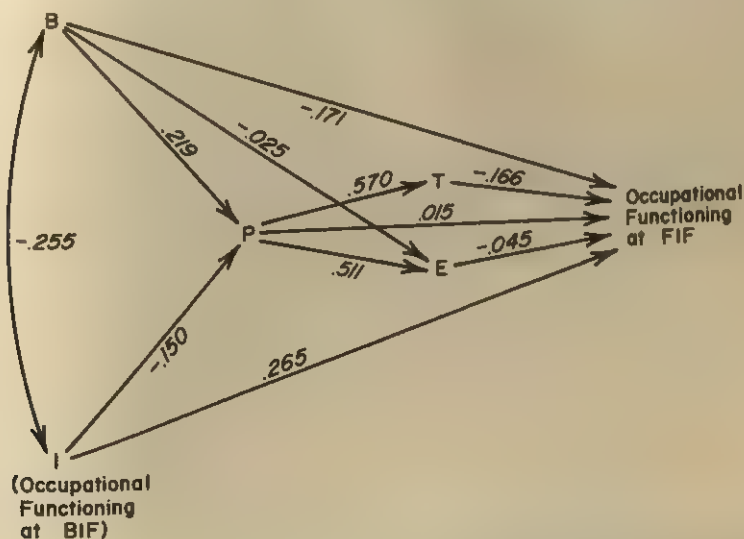


Figure 5. Estimation of path coefficients for occupational functioning. (FIF = follow-up information form; BIF = background information form; B = background; I = intake; P = program; T = treatment experiences; E = treatment environment.)

trast, most of the total effect of the intake symptoms is direct (77%–93%) and thus not shared with the program-related variables.

There is a strong total effect of program type on outcome (q_{63}). When the treatment experiences and perceptions of the environment are taken into account, however, the program variable has little direct effect (p_{63}) (i.e., almost all of the total effect of the program is mediated by the two program-related

variables, shown by the indirect effects, $p_{64}p_{43}$, and $p_{65}p_{53}$, in Table 2).

Treatment experiences (p_{64}) and perceptions of the environment (p_{65}) are both strongly associated with outcome. For all of the outcome criteria except occupational functioning, the direct effect of the perceptions of the environment (p_{65}) is surprisingly strong, suggesting that perceptions of the treatment program may be important predic-

Table 1

Direct, Indirect, and Total Effects of Patient-Related Block Variables on Program-Related Block Variables

Dependent	Variable	Type of intake and outcome criterion used			
		Alcohol consumption	Rating of drinking problem	Physical concomitants	Occupational functioning
Program	Social background				
	direct & total effect: p_{31} (q_{31}) (no indirect effects)	.559*	.522*	.463*	.219*
Treatment experiences	Corresponding intake symptom				
	Direct & total effect: p_{32} (q_{32}) (no indirect effects)	.015	-.080	-.150*	-.150*
	Social background				
	Indirect effect via program $p_{41}p_{31}$ (also total effect—no direct effect) (q_{41})	.257	.360	.274	.124
	Intake symptom				
	Indirect effect via program: $p_{42}p_{32}$ (also total effect—no direct effect) (q_{42})	.007	-.055	-.089	-.086
Perceptions of environment	Direct effect p_{43}	.460*	.689*	.592*	.570*
	Social Background				
	Direct effect: p_{51}	.185*	.104	.122*	-.025
	Indirect effect via program: $p_{52}p_{31}$.243	.269	.256	.112
	Total effect: q_{51} ($p_{51} + p_{52}p_{31}$)	.428	.373	.378	.087
	Intake symptom				
	Indirect effect via program: $p_{53}p_{32}$ (also total effect, q_{52}) (no direct effect)	.007	-.041	-.083	-.077
	Direct effect: p_{53}	.434*	.515*	.552*	.511*

* $p < .05$ (calculated for direct effects only).

Table 2

Direct, Indirect, and Total Effects of Patient-Related Variables and Program-Related Block Variables on Four Outcome Criteria

Effects	Alcohol consumption	Rating of drinking problem	Physical concomitants	Occupational functioning
Direct				
Social background— p_{81} (B)	-.167*	-.061	-.169*	-.171*
Corresponding intake symptom— p_{62} (I)	.036	.171*	.235*	.265*
Program— p_{63} (P)	.068	-.014	.086	.015
Treatment experiences— p_{64} (T)	-.295*	-.155*	-.165*	-.166*
Perceptions of environment— p_{65} (E)	-.190*	-.283*	-.264*	-.045
Indirect				
B via				
P— $p_{63}p_{81}$.038	-.007	.040	.003
E— $p_{65}p_{61}$	-.035	-.029	-.032	.001
P&T— $p_{64}p_{43}p_{81}$	-.076	-.056	-.045	-.021
P&E— $p_{65}p_{63}p_{81}$	-.046	-.076	-.067	-.005
I via				
P— $p_{63}p_{32}$.001	.001	-.013	-.002
P&T— $p_{64}p_{43}p_{32}$	-.002	.009	.015	.014
P&E— $p_{65}p_{63}p_{32}$	-.001	.012	.022	.003
P via				
T— $p_{64}p_{43}$	-.136	-.107	-.098	-.095
E— $p_{65}p_{63}$	-.082	-.146	-.146	-.023
Total*				
B— q_{61}	-.286	-.229	-.273	-.193
I— q_{62}	.047	.193	.259	.284
P— q_{63}	-.150	-.267	-.158	-.103

* Within the context of the causal model used here (see Figure 1), the total effects of treatment experiences and perceptions of the environment, q_{64} and q_{65} , respectively, are the same as their direct effects, p_{64} and p_{65} . Consequently, they are only listed under the direct effects. The total effects listed here were calculated as the sum of the direct and indirect effects.

* $p < .05$ (calculated for direct effects only).

tors of outcome. However, there is substantial interdependence among the program-related variables; only 0%–27% of the direct effect of treatment experiences and 1%–33% of the direct effect of perceptions of the environment are independent of all prior variables (social background, intake symptoms, and program type).⁷

Unique and Joint Contributions of the Explained Variance

Table 3 presents the results of partitioning the explained variance for the four outcome criteria. The first five rows display the unique and shared variance attributed to the patient background variables, and the next six rows show the unique and shared vari-

ance attributed to the program-related variables. The rest of the table displays the explained variance that is shared among combinations of the patient-related and program-related variables.

The pattern of results is similar to those obtained from the path analysis. In general, both social background and the intake symptoms contribute substantially to the explained variance, with the combined contribution of their unique and shared variance ranging from 12% to 61% of the explained variance. In three of the models, the unique contributions

⁷ These percentages were calculated by subtracting out all other paths that are causally prior to the direct path of interest. (See Coleman, 1975, for a more detailed discussion of these procedures.)

Table 3

Partitioning of Explained Variance for Four Outcome Criteria

Variable	Alcohol consumption	Rating of drinking problem	Physical concomitants	Occupational functioning
Patient related				
Unique variance				
Background	.019 ^a	.003	.019 ^a	.026 ^a
Intake	.001	.028 ^a	.047 ^a	.064 ^a
Shared variance				
Background and intake	.002	.003	.019 ^a	.022 ^a
Subtotal	.022 ^a	.034 ^a	.085 ^a	.112 ^a
Proportion of R^2	.12	.16	.32	.61
Program related				
Unique variance				
Program	.002	.000	.003	.000
Treatment experiences	.067 ^a	.012 ^a	.017 ^a	.018 ^a
Perceptions of treatment environment	.025 ^a	.048 ^a	.043 ^a	.002
Shared variance				
Sum of all combinations ^b	.010 ^a	.047 ^a	.013 ^a	.009
Subtotal	.104 ^a	.107 ^a	.076 ^a	.030 ^a
Proportion of R^2	.57	.49	.28	.16
Shared variance among background, intake, & program-related variables				
Sum of all combinations ^c	.056 ^a	.075 ^a	.106 ^a	.041 ^a
Proportion of R^2	.30	.35	.40	.23
Total R^2	.182	.217	.268	.183

Note. B = background; I = intake; P = program; T = treatment experiences; and E = perceptions of treatment environment.

^a The explained variance is greater than 1%.

^b These combinations are PT, PE, TE, PTE.

^c These combinations are BP, BT, BE, BPT, BTE, BPE, BPTE, IP, IT, IE, IPT, ITE, IPE, IPTE, BIP, BIT, BIE, BIPT, BIPE, BITE, and BIPTE.

of the intake symptoms are stronger than the unique explained variance attributed to social background. When taken together with the path analyses, the findings suggest that the intake symptom has a relatively stronger relationship to outcome than does social background.

The total of the unique and shared variance attributed to the program-related variables ranges from 16% to 57% of the explained variance. The unique variance accounted for by the program type is very small, whereas that accounted for by the treatment experiences and perceptions of the environment is larger. This implies that most of the unique variance accounted for by the program-related variables is due to either treatment experiences or perceptions of the environment. For two outcome criteria (rating of drinking problem and physical concomitants), percep-

tions of the environment account for more explained variance, whereas treatment experiences have more explanatory power for the other two (occupational functioning and alcohol consumption). Consistent with the results of the path analyses, the unique contribution of perceptions of the environment is almost as strong or stronger than any of the patient background variables (in three models) or any of the other variables in the model (in two models).

Since each program determines its treatment experiences and the environment it creates, some of the explained variance is shared among combinations of the program-related variables. This corresponds to the path analysis, which shows that most of the total effect of the program type is due to its indirect effects that are shared with the other two program-related variables.

In comparing the contribution of the program-related variables with that of the patient background variables, there is no group of block variables that clearly has the most explanatory power across all models. Patient-related variables account for more explained variance than program-related variables when predicting occupational functioning, whereas their relative contributions are approximately equal when predicting physical concomitants. In contrast, the explained variance attributed to program-related variables is greater in the models of rating of drinking problem and alcohol consumption.

Discussion

We have used an integrated approach to examine the interrelationships among five major sets of variables that are related to treatment outcome. This was done by (a) formulating a model that explicitly specified the hypothesized causal ordering among these sets of variables; (b) constructing block variables to summarize the effects of sets of conceptually similar variables; (c) decomposing the total effects of each block variable into its direct and indirect effects; and (d) partitioning the explained variance into the unique contributions of each block variable and the shared contributions among combinations of block variables. The findings clarify some important issues concerning the way in which patient-related and program-related variables contribute to explaining outcome.

Consistent with previous research (Armor et al., 1976; Bromet et al., 1977; Craft et al., 1975; Ruggels et al., 1975), the results show that social background and intake symptoms are relatively strong predictors of outcome. In addition, although most of the total effect of the intake symptoms is direct, social background has substantial indirect effects that are mediated by the program-related variables. These effects represent shared variance between sociodemographic characteristics and program-related variables and cannot be attributed solely to either set of variables alone.

Although the relative importance of social background and intake symptoms is not consistent across all four models, the intake symptoms have stronger direct effects and

account for a larger proportion of the explained variance for three of the four outcome criteria. Even though no clear-cut pattern has been reported in previous research, these findings suggest that the intake symptoms may be slightly stronger predictors of some outcome criteria than social background. More research on different types of patient functioning characteristics is needed to clarify this issue.

Previous research has reported that treatment does contribute to improved functioning from intake to follow-up; that is, recovery rates of treated alcoholic patients vary directly with the amount of treatment, length of stay, and level of participation in program activities (Armor et al., 1976; Bromet et al., 1977; Craft et al., 1975). Although the total effect of program type on outcome is relatively strong in the present analyses, it is primarily indirect and accounted for through the effects of treatment experiences and perceptions of the environment. The strong relationship between these program-related variables and outcome may reflect not only treatment effects but also the combined effects of the patient's motivation to recover, a more positive attitude toward the program, better functioning within the program, and a greater probability of participating in aftercare services (Craft et al., 1975; Pisani, 1969; Pratt, Linn, Carmichael, & Webb, 1977).

In comparing the importance of patient characteristics relative to program-related variables as predictors of outcome, previous studies have reported relative uniformity in outcome among programs when patient-related variables are taken into account (Armor et al., 1976; Ruggels et al., 1975). Contrary to previous findings, the combined unique effects of the program-related variables observed here are substantial. In fact, except for occupational functioning, the proportion of explained variance uniquely accounted for by program-related variables is almost equal to or greater than that uniquely accounted for by patient-related variables. These findings indicate that program-related effects may be more important than would be expected from previous research.

The results also show that 23%–40% of the total explained variance is shared between patient-related and program-related variables (see Table 3). From a different perspective, between 28% and 72% of the total effects of the patient characteristics are shared with the program-related variables. In previous research, the patient-related variables have been credited with these shared effects, leading researchers to underestimate the explanatory power of program-related variables. Given that a substantial portion of the variance is shared between these two groups of variables, researchers may have been too limited in their analyses by looking primarily for either patient or program variance. The effects of certain combinations of patient and program-related variables, such as patient-program selection and congruence effects, may be particularly important for understanding patterns of patient improvement (Pattison, 1976).

Although there are some consistencies across all four models, there are also some important differences. For example, although the intake symptom is one of the two most important predictors for three outcome criteria, it is least important in the alcohol consumption model. Apparently, the level of alcohol consumption is more strongly influenced by treatment experiences, perceptions of the treatment environment, and social background. In contrast, occupational functioning at follow-up is most strongly affected by occupational functioning at intake and much less influenced by the program-related variables. These variations indicate that the importance of program-related variables relative to background variables depends on the outcome criterion used.

Consistent with previous research on program selection, our findings show that social background is an important determinant of the type of program that a patient enters (Armor et al., 1976; Bromet et al., 1977). In addition, the findings suggest that for two patient functioning characteristics (occupational functioning and physical concomitants), the intake symptoms are also a significant determinant of program selection. These two intake symptoms are those most highly correlated with social background, and

their effects may thus reflect some underlying effects of social background.

The program is by far the most important determinant of a patient's treatment experiences and perceptions of the environment. Although the overall effect of social background on perceptions of the environment is positive here, other findings (Moos & Bromet, 1978) suggest that the effect is relatively small, and that the effects of different background variables may vary (e.g., women may perceive the environment more positively, whereas better educated patients may perceive it more negatively). Social background also has substantial indirect effects, via the program, on both treatment experiences and perceptions of the environment. These indirect effects suggest that patients with higher social background levels either enter and/or participate more actively in programs offering environmental and treatment experiences associated with better outcome. In addition, the indirect effects may reflect the fact that patients with certain background characteristics are (a) more motivated to become involved in program activities and thus receive more treatment and (b) function better within the program and thus perceive the environment as more positive.

One of the most distinctive features of the present analysis is the use of block variables. The method of determining weights that we used is one of several approaches that can be used. Alternative methods include giving the variables equal weights by averaging them, giving one or more variables a weight of one and others a weight of zero by dropping variables from the set, and using various types of factor analyses. The important point is that by grouping together a set of conceptually similar variables that are then treated as one, subsequent analyses provide a more integrated and comprehensive picture of a complicated pattern of effects.

Between 18% and 27% of the overall variance in treatment outcome is explained by all of the block variables taken together. Although this is a somewhat larger proportion of the overall variance than that accounted for in most similar studies, the majority of the variance in outcome is still unaccounted for by either patient-related or

program-related variables. Further research is needed to identify other important factors, such as the environmental resources available to patients in community settings (Bromet & Moos, 1977), that may contribute to the remaining variance. Nevertheless, the results presented here indicate that the proportion of variance in the outcome criteria uniquely associated with program-related variables is as great or greater than that uniquely associated with patient-related variables. Perhaps the most important conclusion to be derived from these results is that alcoholism treatment programs may have more substantial differential effects on outcome than previous literature has suggested. Since some of these effects are probably related to patient-program selection and congruence, it may be unwise to implement the policy of placing all patients in uniform low-cost treatment programs.

References

- Alwin, D., & Hauser, R. The decomposition of effects in path analysis. *American Sociological Review*, 1975, 40, 37-47.
- Armor, D., Polich, J., & Stambul, H. *Alcoholism and treatment*. Santa Monica, Calif.: Rand Corporation, 1976.
- Bromet, E., & Moos, R. Environmental resources and the posttreatment functioning of alcoholic patients. *Journal of Health and Social Behavior*, 1977, 18, 326-338.
- Bromet, E., Moos, R., & Bliss, F. The social climate of alcoholism treatment programs. *Archives of General Psychiatry*, 1976, 33, 910-916.
- Bromet, E., Moos, R., Bliss, F., & Wuthmann, C. The posttreatment functioning of alcoholic patients: Its relation to program participation. *Journal of Consulting and Clinical Psychology*, 1977, 45, 829-842.
- Coleman, J. S. Methods and results in the IEA studies of effects of school on learning. *Review of Educational Research*, 1975, 45, 335-386.
- Craft, J., Sheehan, D., Driggers, J., & DuBois, R. *Components of effectiveness in alcoholism treatment programs: V. Correlates of treatment outcomes*. Austin: Texas Department of Mental Health and Mental Retardation, 1975.
- Ellsworth, R., Maroney, R., Klett, W., Gordon, H., & Gunn, R. Milieu characteristics of successful psychiatric treatment programs. *American Journal of Orthopsychiatry*, 1971, 41, 427-441.
- Gerard, D., & Saenger, G. *Out-patient treatment of alcoholism: A study of outcome and its determinants*. Toronto, Canada: University of Toronto Press, 1966.
- Kissin, B., Platz, A., & Su, W. Social and psychological factors in the treatment of chronic alcoholism. *Journal of Psychiatric Research*, 1970, 8, 13-27.
- Mood, A. M. Partitioning variance in multiple regression analyses as a tool for developing learning models. *American Educational Research Journal*, 1971, 8, 191-202.
- Moos, R. *Community-Oriented Programs Environment Scale manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1974. (a)
- Moos, R. *Evaluating treatment environments: A social ecological approach*. New York: Wiley, 1974. (b)
- Moos, R. H., & Bromet, E. Relation of patient attributes to perceptions of the treatment environment. *Journal of Consulting and Clinical Psychology*, 1978, 46, 350-351.
- Newton, R. G., & Spurrell, D. J. A development of multiple regression for the analysis of routine data. *Applied Statistics*, 1967, 16, 51-64. (a)
- Newton, R. G., & Spurrell, D. J. Examples of the use of elements for clarifying regression analyses. *Applied Statistics*, 1967, 16, 165-172. (b)
- Pattison, E. M. A conceptual approach to alcoholism treatment goals. *Addictive Behaviors*, 1976, 1, 177-192.
- Pisani, V. Assessing inpatient attitudes toward an alcoholism treatment center. *Quarterly Journal of Studies on Alcohol*, 1969, 30, 640-644.
- Pokorny, A., Miller, B., & Cleveland, S. Response to treatment of alcoholism: A follow-up study. *Quarterly Journal of Studies on Alcohol*, 1968, 29, 364-381.
- Pratt, R., Linn, M., Carmichael, J., & Webb, N. The alcoholic's perception of the ward as a predictor of after care attendance. *Journal of Clinical Psychology*, 1977, 33, 915-918.
- Ruggels, W., Armor, D., Polich, J., Mothershead, A., & Stephen, M. *A follow-up study of clients at selected alcoholism treatment centers funded by NIAAA*. Menlo Park, Calif.: Stanford Research Institute, 1975.
- Sobell, L. C., & Sobell, M. B. Outpatient alcoholics give valid self-reports. *Journal of Nervous and Mental Disease*, 1975, 161, 32-42.

Received September 26, 1977 ■

Therapist Experience, General Clinical Ability, and Treatment Outcome in Schizophrenia

A. Hussain Tuma
Clinical Research Branch,
National Institute of Mental Health
Rockville, Maryland

Coralee Yale
Health Sciences Computing Facility
University of California, Los Angeles

Philip R. A. May
Health Services and Research Development
Laboratory, Veterans Administration Hospital,
Brentwood, California, and Neuropsychiatric
Institute, University of California

Alan B. Forsythe
Department of Biomathematics and Biostatistics
University of California, Los Angeles

Knowledge of the effect of therapist training and experience on the outcome of treatment of schizophrenic patients is scanty. This article presents data systematically collected in the course of a controlled comparison of the effects of five different treatment methods in schizophrenia. Among the 23 outcome variables studied, there was not a single instance in which the effect of therapist experience and general clinical ability was significantly related to outcome. There appeared to be, however, differences among therapists' results that were *not* related to experience and general clinical ability, particularly in relation to the length of time that they kept their patients in hospital. Drug treatment tended to override but perhaps not entirely eliminate these effects.

Do Experienced Therapists Get Better Results?

Research evidence on this question is scanty. The vast gap in knowledge is typically filled in with a mixture of myth, speculation, anecdote, and weak research evidence, mostly relating to psychotherapy. (Betz & Whitehorn, 1956; Karon & VandenBos, 1972, 1975; Tuma & May, 1975). The notion that experience and training might influence the results of treat-

ments other than psychotherapy has been largely ignored or depreciated.

This article presents data on this issue that were systematically collected in the course of a controlled comparison of the outcome of five different methods of treatment in schizophrenia.

Background and Literature Review

The literature virtually ignores all methods of treatment except psychotherapy. Even in this area the studies are generally anecdotal or nonexperimental, with small numbers of cases and equivocal or contradictory conclusions. Strupp stated that there is some anecdotal evidence (Glover, cited in Kubie, 1956) that beginners may achieve success in psychotherapy that they are unable to equal once they have had formal training; but on the other hand, he continued, highly experienced therapists might be presumed to achieve better results, perhaps because they are more circumspect in selecting their patients (Strupp, 1958a, 1958b, 1958c, 1960). Experienced therapists are probably better aware of their own strengths and weaknesses and are not likely

The research to which this work is related has been supported in part by research grants from the State of California Department of Mental Hygiene and Grants MH 02719 and MH 04589 and Contract PH 43-66-49 from the National Institute of Mental Health, and by the Research Service of the Veterans Administration. Computing assistance from the Health Sciences Computing Facility at the Center for the Health Sciences, University of California at Los Angeles, was sponsored by National Institutes of Health Grant RR-3.

Jane A. Kirksey provided assistance in data processing and Hilda Margolis and Dona McLemore prepared the manuscript and table.

Requests for reprints should be sent to A. Hussain Tuma, Clinical Research Branch, National Institute of Mental Health, 5600 Fishers Lane, Rockville, Maryland 20852.

to undertake tasks that they believe they cannot tackle. Unfortunately, we have little data on well-trained, highly experienced therapists and the reasons for their presumed superior accomplishments.

An early series of studies by Fiedler (1950a, 1950b, 1951) of experienced and inexperienced therapists' description of their own attitudes and behaviors in psychotherapy is often cited in this context. However, Fiedler did not investigate whether the therapists' self-description has any bearing on actual outcome. Nor were his observations based on the treatment of psychotic patients. Thus, although Fiedler's studies focused attention on what therapists considered important in psychotherapy regardless of theoretical orientation, his data did not shed any light on the influence of therapists' level of experience on the outcome of treatment.

Carkhuff and Truax (1965) and Banks, Berenson, and Carkhuff (1967) studied changes in such characteristics of therapists as empathy, warmth, genuineness, depth of intrapersonal exploration, and positive regard in response to training and experience. Again, although the studies provided evidence that therapists' behavior during psychotherapy may be modifiable, they say nothing about therapists' characteristics relative to treatment outcome.

There is a common belief that inexperienced therapists, being relatively simple and enthusiastic, may get better results particularly with schizophrenic patients, than those who have been disillusioned by experience. Poser (1966) compared trained and untrained therapists conducting group therapy with hospitalized chronic schizophrenic patients. The untrained achieved slightly better results than the trained, and these results persisted during a 3-year follow-up. However, the trained and untrained therapists did not treat their patients at the same time, and there was a high proportion of dropouts (Rosenbaum, 1966). Barrett-Lennard (1962) compared expert counselors with nonexperts. The experts kept their clients in treatment longer and obtained more improvement at the .10 level of significance. Improvement was, however, measured by the therapists' own ratings—not independently. This could reflect the effect of self-deprecatory attitudes by the nonexperts,

enhanced self-esteem among the experts, or even the longer treatment time.

In general, the relatively few systematic though naturalistic studies of this matter (Goldberg, Schooler, Davidson, & Kayce, 1966; Cole, Note 1), point to the conclusion that psychiatrists who have completed their residency training and residents who are still at various stages of their training all obtain similar results.

With regard to drug therapy, it has been suggested that only inexperienced therapists get better results, but the findings are somewhat contradictory. In a study by Rickels et al. (1966), five board-certified or board-eligible psychiatrists treated anxious neurotic patients with psychotherapy plus placebo or with psychotherapy plus meprobamate. The combination with drug was significantly superior by a variety of criteria. Uhlenhuth, Covi, Rickels, Lipman, and Park (1972), however, reported that experienced doctors got better results with placebo, whereas the inexperienced got better results with meprobamate.

In schizophrenia, reports are also conflicting. Karon and O'Grady (1969) and Karon and VandenBos (1970, 1972) reported that patients treated by 2 experienced therapists improved more and spent less time in hospital over a 2-year follow-up than those treated by 10 inexperienced therapists: This applied whether or not drugs were used. It should not be inferred that experienced therapists got better results without drugs than with them. Indeed, well-controlled studies by Grinspoon, Ewalt, and Shader (1967a, 1967b, 1968, 1972) reported that experienced therapists obtained better results in schizophrenia with drugs plus psychotherapy than with psychotherapy alone.

Karon and co-workers also reported that when drugs were used, the inexperienced therapists kept their patients in the hospital for less time at the beginning but with less improvement in thought disorder and greater long-term hospital stay. This is, however, misleading. In fact, the patients treated without drugs spent more time (not less) in the hospital if one considers the entire period from admission to follow-up (see May & Tuma, 1970).

O'Brien et al. (1972) compared the results obtained by medical students, social workers, and psychiatrists using individual and group

psychotherapy combined with drugs for schizophrenic outpatients. There was a tendency for the medical students to get the best results, but the differences between the three groups were not statistically significant.

Clearly, there is a lack of relevant and unambiguous data in this area and a distinct need for careful studies.

Method

The data were collected in the context of a larger study in which the primary aim was to evaluate the differential effectiveness of five methods currently used in the treatment of schizophrenia. All methods are still in use today, with little, if any, difference in procedure and technique from those used then. The experimental design and procedures are described elsewhere in considerable detail (Dixon & May, 1968; May, 1968); accordingly, only certain details that are specific to the topic of this study will be presented.

Two hundred twenty-eight male and female first-admission schizophrenic patients without significant prior treatment were assigned by a stratified random method to five treatment groups: (a) individual psychotherapy alone; (b) ataractic drug alone; (c) individual psychotherapy plus drug; (d) electroconvulsive therapy; and (e) milieu, a group that received none of the above "specific" treatments but only the same level of basic milieu care given to all the other groups. The number of patients in each group was as follows: psychotherapy, 42 patients; drug, 48; psychotherapy plus drug, 44; electroconvulsive therapy, 47; and milieu, 43.

The therapists were 33 male and 5 female psychiatric residents or recently (1-3 years) graduated psychiatrists who treated an average of six patients each. Patients were assigned to their therapists on the basis of openings on a rotating roster, and each patient's treatment was supervised by a senior consultant experienced in the particular treatment to which the patient was assigned. Unless released from hospital earlier, patients were treated for a minimum of 6 months to up to 1 year before treatment was terminated.

Our analyses include data relevant to all treatment groups. It could be argued that the therapist is more central or critical when formal psychotherapy is undertaken. This is certainly a legitimate point of view, and specific analyses might have been undertaken to examine the question in this limited context. However, this study is equally concerned with the role of therapists' experience and ability in all five treatments. The core hypothesis is that therapists' experience and clinical ability may influence many of their behaviors, judgments, and decisions, which may in turn influence outcome.

Measures of Outcome

The 23 outcome measures chosen for this analysis are presented in Table 1. Additional details can be found elsewhere (May, 1968).

Table 1
Measures of Outcome

Variable	Rater
Menninger Health-Sickness Scale, post-treatment	2 independent raters (psychoanalysts)
Camarillo Dynamic Assessment Scale ^a	2 independent raters (psychoanalysts)
MMPI, posttreatment <i>F, Sc, Pa</i> , psychiatric triad	Patient
Testability, posttreatment	Psychologist
MACC Communication, Total score	Nurses
Hospital stay	
Only patients successfully discharged	—
All patients	—
Release rate ^b	—
Clyde Mood Scale, post-treatment (clear thinking scale)	Therapist
Symptom rating sheet, posttreatment, total score	Therapist

Note. MMPI = Minnesota Multiphasic Personality Inventory.

^a Consists of eight scales: Affective Contact; Anxiety Level; Ego Strength; Extent to which Environment Suffers; Insight; Motivation; Object Relations; Sense of Personal Identity.

^b Counted when patient is successful in staying out of the hospital 31 days or more.

Therapist Variables

Length of experience. A number of indices of experience were considered. These included age, number of years since internship, number of years in psychiatric residency training, and number of years of clinical experience in treating psychiatric patients.

Since treatment lasted a relatively long period of time ($M = 169.4$ days), and since the treatment phase of the study lasted several years, the indices of therapist experience were computed using the midpoint of each patients' treatment. This allows for the fact that therapists may be more or less experienced when they treat particular patients. Data describing the therapist sample on all four indices of experience were analyzed, and there were no significant differences among the treatment groups with respect to any item.

The number of years of *clinical experience in treating psychiatric patients* was chosen as the most appropriate index of "experience" for the present analysis, even

though the other three indices are also potentially relevant.¹ The length of this experience ranged from .16 to 11.65 years ($M = 3.77$, $SD = 2.87$). There was no significant difference² among treatment groups ($p = .1664$ for all patients; and $p = .3540$ for those with only one therapist). A two-way analysis of variance for the milieu, psychotherapy alone, psychotherapy plus drug, and drug-alone groups examined the differences in therapist psychiatric experience between the two groups that received psychotherapy and those that did not receive psychotherapy. This resulted in a p value of .0281 for all patients, and $p = .0334$ for those with only one therapist. This reflects a deliberate design bias in favor of psychotherapy, in that, as pointed out in an earlier publication (May 1968), therapists with less than 6 months of experience were not assigned to treat patients in psychotherapy.

Therapists' general clinical ability. Every 6 months a rating was made by a senior psychiatrist of the therapist's general clinical ability relative to others of a comparable level of training. The rating made on the date closest to the midpoint of each patient's treatment was used in the present analysis. No significant differences were found between the treatment groups in terms of ratings of therapists' general clinical ability ($p = .3006$ for all patients, and $p = .2303$ for those with only one therapist).

Patient Variables

Seven patient covariates were used to adjust for initial pretreatment differences among patients: presence or absence of precipitating stress; intensity of such stress relative to onset of illness; duration of psychotic disorder; nurses' pretreatment ratings of cooperation on the Movement, Affect, Communication, and Cooperation (MACC) scale (Ellsworth, 1957; Ellsworth & Clayton, 1959); psychoanalysts' pretreatment ratings of affective contact, best level ever attained (May & Dixon, 1969); age at onset of any symptoms; and pretreatment level on the outcome measure under study. Ratings on the Menninger Health-Sickness Scale (Luborsky, 1962) were used to adjust hospital stay and release rate, since no initial level on these measures is obviously available or applicable.

Statistical Analysis

General. An analysis of covariance was performed in which the criterion score of any particular patient was taken as comprising portions of variation associated with: (a) therapist's years of experience, (b) therapist's general clinical ability relative to peers, (c) drug treatment, (d) pretreatment patient variables, and (e) other unknown parameters (error).

Covariance runs. The covariates (i.e., patient, therapist, and drug covariates) were examined separately, first for all patients and then for patients with one therapist (the majority of cases). Wherever there were two therapists, the data relate to the patient's major therapist, that is, the one who treated the patient for the longest period of time.

Interpretation. The analyses examined whether there were differences in outcome among therapists and to what degree these differences could be attributed to the specified covariates.

1. Are the differences in outcome attributable to a particular covariate or group of covariates; that is, is the regression slope significantly different from zero?

2. Are there differences in outcome among therapists before and after adjusting out the covariate; that is, are the therapist cell means significantly different?

The probability of residual differences among therapists after adjusting out the covariate is compared with the probability before covariance adjustment.

Results

Was Therapist Experience Significantly Related to Treatment Outcome?

Among the 23 outcome variables, there was not a single instance in which the regression slopes for the effects of therapist experience and general clinical ability were significantly different from zero, whether for all cases or for those with only one therapist.³ Indeed, the relationship reached borderline significance for only one outcome variable, length of stay for patients successfully released only, not including the failures ($p = .0875$). This relationship was nonsignificant ($p = .6123$) when examined for cases with only one therapist. Our inclination is to discount this lone marginal finding as a chance occurrence in the midst of a great number of insignificant p values.

Were There Significant Differences in Outcome Attributable to Therapist Factors Other Than Experience and General Clinical Ability?

Length of stay. After adjusting for the effects of therapist experience and general clinical ability, there were significant differences among therapists in the length of time that they kept their patients in hospital. The adjusted therapist means were significantly different for length of hospital stay for patients who were successfully released only, not including the failures ($p = .0131$ for all cases,

¹ Data on all indices can be made available to investigators on request.

² In this article the term *significant* is used if $p < .05$ and *borderline significance* if $.10 > p > .0501$.

³ To save space no detailed tables will be presented here, only p values will be presented as appropriate.

and $p = .0239$ for cases with only one therapist). For length of stay for all patients, including the failures, $p = .0927$ for all cases and $p = .0362$ for cases with only one therapist. These differences in patients' hospital stay could not be attributed to their therapist's experience in the treatment of mental illness, to ratings of their general clinical ability, or to the seven patient covariates examined, since the differences remained significant even after adjusting for patient differences on these covariates. (After adjustment for length of stay to successful release, $p = .0134$ for all cases, and $p = .0234$ for cases with only one therapist. For length of stay including the failures, $p = .0733$ for all cases, and $p = .0369$ for those with only one therapist.)

There was some evidence that these therapist-related differences in length of hospital stay were partly, but not entirely, related to whether or not the patient was given drug treatment, since the significance levels were sharply reduced when further adjustment was made for drug treatment. (For only the patients who were released, the significance level changed from .0134 to .0837 for all cases; p changed from .0234 to .1662 for those with only one therapist. For all cases including the failures, the p levels changed from .0927 to .2263 and from .0363 to .1534, respectively.)

Other outcome measures. After adjusting for the effect of therapist experience and clinical ability, there were significant or borderline differences among therapists' results for only 1 of the other 21 outcome measures, the Minnesota Multiphasic Personality Inventory *Pa* scale (Dahlstrom & Welsh, 1960; Hathaway & McKinley, 1943) ($p = .0467$ for all cases, and $p = .0349$ for those with only one therapist). This should probably be dismissed as chance occurrence among so many analyses.

Again, the significance of the difference is reduced if further adjustment is made for whether or not drug treatment was given ($p = .1546$ for all cases, and $p = .1458$ for those with only one therapist).

Discussion

These results are presented with some hesitation for we are mindful that some may take them as definitive findings from an experiment

in which both therapists and patients were randomly assigned to a given treatment, which they were not; as applying to all types of mental disorder besides schizophrenia, which they do not; and as applying to the (small and select number of) highly experienced persons who specialize in the treatment of schizophrenia, which they do not. They do deserve some credence, however, as an epidemiologic report from a carefully controlled study with a moderate but carefully defined and measured range of therapist experience, in which patients were randomly assigned to treatment and therapists were assigned from a rotating roster, with no self-selection of patients.

Within these limitations, some of the findings encourage further work on the relationship between therapist characteristics and the outcome of their treatment of schizophrenic patients. The findings suggest that there are possible differences among therapists in the results they obtain. Such differences may, however, bear little relationship to therapist experience or general clinical ability (as defined and within the range sampled in this study) or to the usual patient prognostic factors. They may turn out to be related more to other therapist characteristics, perhaps of personality, sophistication in the use of various treatment strategies and activities, and ability to handle paranoid or hostile attitudes.

The point must be made, however, that these findings do not necessarily mean that training and experience make no difference at all. Obviously an inexperienced and unsupervised or inept therapist could abuse or misuse any form of treatment, whereas a talented, versatile, and highly experienced therapist could optimize its value. A more likely interpretation of our data is that the kind of intensive supervision that our therapists received in addition to the usual teamwork approach to treatment in general may have been sufficient to compensate for any differences in results that might otherwise have occurred due to therapists' factors.

In future work, the effect of drug treatment must be carefully taken into account, since this powerful form of treatment seems to override or compensate for differences in outcome that might otherwise occur among therapists.

It is concluded that future research in this area should focus on identifying *specific* therapist characteristics other than general clinical experience and general clinical ability (as defined in this study) that relate to *specific* outcome characteristics. In this way progress might be made toward eventually matching therapist and patients *when a specific outcome characteristic is desired*. Perhaps, therefore, the research questions should be rephrased in more sophisticated form: At what point for each particular type of clinical problem and for each particular type of treatment do training and experience begin to influence treatment results, and is there a point of diminishing returns? What is the shape of the experience/effectiveness curve? Is there an optimum cost-effective point in training and experience?

It may be that the future will bring rigorous experiments designed to answer such questions. At this time, however, we are forced to rely on less definitive material.

Reference Note

1. Cole J. O. Personal communication, December 12, 1967.

References

- Banks, G., Berenson, B. G., & Carkhuff, R. R. The effects of counselor race and training upon counseling process with Negro clients in initial interviews. *Journal of Clinical Psychology*, 1967, 23, 70-72.
- Barrett-Lennard, G. T. Dimensions of therapist response as causal factors in therapeutic change. *Psychological Monographs*, 1962, 76(43, Whole No. 562).
- Betz, B. J., & Whitehorn, J. C. The relationship of the therapist to the outcome of therapy in schizophrenia. *Psychiatric Research Reports*, 1956, 5, 89-117.
- Carkhuff, R. R., & Truax, C. B. Training in counseling and psychotherapy: An evaluation of an integrated and experiential approach. *Journal of Consulting Psychology*, 1965, 29, 333-336.
- Dahlstrom, W. G., & Welsh, G. S. Scales and scoring procedures. In, *An MMPI handbook*. Minneapolis: University of Minnesota Press, 1960.
- Dixon, W. J., & May, P. R. A. Methods of statistical analysis. In P. R. A. May, *Treatment of schizophrenia*. New York: Science House, 1968.
- Ellsworth, R. B. The MACC Behavioral Adjustment Scale. Santa Monica, Calif.: Western Psychological Services, 1957.
- Ellsworth, R. B., & Clayton, W. H. Measurement of improvement in mental illness. *Journal of Consulting Psychology*, 1959, 23, 15-20.
- Fiedler, F. E. A comparison of therapeutic relationships in psychoanalytic, non-directive and Adlerian theory. *Journal of Consulting Psychology*, 1950, 14, 436-445. (a)
- Fiedler, F. E. The concept of an ideal therapeutic relationship. *Journal of Consulting Psychology*, 1950, 14, 239-245. (b)
- Fiedler, F. E. Factor analysis of psychoanalytic, non-directive and Adlerian therapeutic relationships. *Journal of Consulting Psychology*, 1951, 15, 32-38.
- Goldberg, S. C., Schooler, N. R., Davidson, E. M., & Kayce, M. M. Sex and race differences in response to drug treatment among schizophrenics. *Psychopharmacologia*, 1966, 9, 31-47.
- Grinspoon, L., Ewalt, J. R., & Shader, R. I. Long term treatment of chronic schizophrenia: A preliminary report. *International Journal of Psychiatry*, 1967, 4, 116-128. (a)
- Grinspoon, L., Ewalt, J. R., & Shader, R. I. A preliminary report on long term treatment of chronic schizophrenia. *International Journal of Psychiatry*, 1967, 4, 116-128. (b)
- Grinspoon, L., Ewalt, J. R., & Shader, R. I. Psychotherapy and pharmacotherapy in chronic schizophrenia. *American Journal of Psychiatry*, 1968, 124, 1645-1652.
- Grinspoon, L., Ewalt, J. R., & Shader, R. I. *Schizophrenia, pharmacotherapy and psychotherapy*. Baltimore, Md.: Williams & Wilkins, 1972.
- Hathaway, S. R., & McKinley, J. C. *The Minnesota Multiphasic Personality Inventory*. Minneapolis: University of Minnesota Press, 1943.
- Karon, B. P., & O'Grady, P. Intellectual test changes in schizophrenic patients in the first six months of treatment. *Psychotherapy: Theory, Research and Practice*, 1969, 6, 88-96.
- Karon, B. P., & VandenBos, G. R. Experience, medication, and the effectiveness of psychotherapy with schizophrenics: A note on Drs. May and Tuma's conclusions. *British Journal of Psychiatry*, 1970, 116, 427-428.
- Karon, B. P., & VandenBos, G. R. The consequences of psychotherapy for schizophrenic patients. *Psychotherapy: Theory, Research and Practice*, 1972, 9, 111-119.
- Karon, B. P., & VandenBos, G. R. Issues in current research on psychotherapy vs. medication in treatment of schizophrenia. *Psychotherapy: Theory, Research and Practice*, 1975, 12, 143-148.
- Kubie, L. S. Some unsolved problems of psychoanalytic psychotherapy. In F. Fromm-Reichmann & J. L. Moreno (Eds.), *Progress in psychotherapy*. New York: Grune & Stratton, 1956.
- Luborsky, L. Clinician's judgments of mental health: A proposed scale. *Archives of General Psychiatry*, 1962, 7, 407-417.
- May, P. R. A. *Treatment of schizophrenia*. New York: Science House, 1968.
- May, P. R. A., & Dixon, W. J. The Camarillo Dynamic Assessment Scales. *Bulletin of the Menninger Clinic*, 1969, 33, 1-35.
- May, P. R. A., & Tuma, A. H. Methodological problems in psychotherapy research: Observations on the Karon-VandenBos study of psychotherapy and

- drugs in schizophrenia. *British Journal of Psychiatry*, 1970, 117, 569-570.
- O'Brien, C. P., et al. Group versus individual psychotherapy with schizophrenics: A controlled outcome study. *Archives of General Psychiatry*, 1972, 27, 474-478.
- Poser, E. G. The effect of therapists' training on group therapeutic outcome. *Journal of Consulting Psychology*, 1966, 30, 283-289.
- Rickels, R., et al. Controlled psychopharmacological research in private psychiatric practice. *Psychopharmacologia*, 1966, 9, 288-306.
- Rosenbaum, M. Some comments on the use of untrained therapists. *Journal of Consulting Psychology*, 1966, 30, 292-294.
- Strupp, H. H. The performance of psychiatrists and psychologists in a therapeutic interview. *Journal of Clinical Psychology*, 1958, 14, 219-226. (a)
- Strupp, H. H. The performance of psychoanalytic and client-centered therapists in an initial interview. *Journal of Consulting Psychology*, 1958, 22, 265-274. (b)
- Strupp, H. H. The psychotherapist's contribution to the treatment process: An experimental investigation. *Behavioral Science*, 1958, 3, 34-67. (c)
- Strupp, H. H. Nature of psychotherapists' contribution to treatment process. *Archives of General Psychiatry*, 1960, 3, 219-321.
- Tuma, A. H., & May, P. R. A. Psychotherapy, drug, and therapist experience in the treatment of schizophrenia. *Psychotherapy: Theory, Research and Practice*, 1975, 12, 138-142.
- Uhlenhuth, E. H., Covi, L., Rickels, K., Lipman, R. S., & Park, L. C. Predicting the relief of anxiety with meprobamate. *Archives of General Psychiatry*, 1972, 26, 85-92.

Received October 3, 1977

An MMPI Scale to Separate Brain-Damaged From Functional Psychiatric Patients in Neuropsychiatric Settings

Charles G. Watson and Duane Plemel
Veterans Administration Hospital, St. Cloud, Minnesota

An empirical Minnesota Multiphasic Personality Inventory scale was developed to separate brain-damaged from functional psychiatric patients. It consisted of 56 items, which significantly differentiated organic and functional groups in a psychiatric hospital and was named the Psychiatric-Organic (*P-O*) scale. Upon cross-validation it was found capable of separating organics from process and reactive schizophrenics, alcoholics, and neurotics, as well as patients with character disorders and affective psychoses. Additionally it was found that by using the scale in combination with a traditional brain-damage test (the Benton Visual Retention Test), better discrimination could be achieved than was possible with either measure alone. The comparative probabilities of functional and organic diagnoses for various *P-O* scale ranges are presented.

The literature on the separation of brain-damaged from nonorganic psychiatric patients via measures of ability is generally unencouraging. Attempts to differentiate organics from schizophrenics with such instruments as the Halstead battery (Watson, Thomas, Andersen, & Felling, 1968); the Trail-Making Test (Brown, Casey, Fisch, & Neuringer, 1958); the Critical Flicker Fusion Test (Watson, Thomas, Felling, & Andersen, 1969); the Bender-Gestalt; and the Graham-Kendall Memory-For-Designs Test (Watson, 1968) have often met with failure. Somewhat more success has been achieved by investigators using personality measures to separate the two groups (Russell, 1975; Watson, 1971; Watson & Thomas, 1968). The Minnesota Multiphasic Personality Inventory (MMPI) Schizophrenia-Organicity (*Sc-O*) scale, which was developed for that purpose (Watson, 1971), has proven to be of value and has been cross-

validated successfully in at least six settings (Ayers, Templer, & Ruff, 1975; Holland, Lowenfeld, & Wadsworth, 1975; Watson, 1971; Andersen, Felling, & Seitz, Note 1; Gilbertstadt, Note 2). In addition, personality and ability measures have been combined to enhance separation of organic and schizophrenic patients (Watson, 1973).

There is also reason to believe that the separation of organics from other functional disorders is problematic as well. Watson, Davis, and Gasser (1978) found that such commonly used instruments as the Halstead Category; Benton Visual Retention; Smith Digit Modality; and Wechsler Adult Intelligence Scale (WAIS) Digit Span, Block Design, and Object Assembly tests are of little value in the differentiation of psychiatric hospital organics from depressives. These findings suggest that research designed to develop a personality scale capable of separating organics from individuals suffering from all types of functional disorders is needed. Although one might have hoped that the *Sc-O* would separate nonschizophrenic functional groups from organics, Watson's (1973) article indicates that it does not, and that a new scale is needed. Accordingly, the research program described here was formulated to develop an MMPI scale for that purpose.

This research was supported by the Veterans Administration Medical Research Service. The authors are indebted to Betty Gasser, Allan Schaefer, Lyle Jacobs, Teresa Kucala, Gail Bamber, and Ed Medeck for their contributions to this research.

Requests for reprints should be sent to Charles G. Watson, Research Service, Veterans Administration Hospital, St. Cloud, Minnesota 56301.

Table 1
Mean Age, Education, and Estimated Full
Scale IQ in Scale-Construction Groups

Variable	Brain damaged	Psychiatric	t
Age	48.3	40.0	3.71*
Education	10.1	11.3	2.66*
Estimated IQ	96.0	109.0	4.75*

* $p < .01$.

Method

Scale Construction

Patients referred to the Psychology Service at the St. Cloud (Minnesota) Veterans Administration Hospital specifically for evaluation of possible brain damage served as subjects for item selection. The brain-damaged group consisted of 40 individuals who were eventually given diagnoses of organic brain syndrome, and whose physician, ward nurse, and psychologist all agreed on the basis of history, laboratory, and/or clinical findings that brain damage was present. (The raters were instructed to make their assessments independent of psychological test data.) The controls were 60 patients who had received functional diagnoses and whose ward physician, psychologist, and nurse agreed that clinically detectable brain damage was not present. Only patients under 60 years of age were included. When more than one testing was available, the MMPI scores taken at the time point closest to referral were used. Only potential subjects with MMPIs taken within 1 month of referral were included.

As might have been expected, the groups differed significantly on mean age, education, and Henmon-Nelson-based estimate (Watson & Klett, 1975) of WAIS Full Scale IQ. These means are reported in Table 1. According to the patients' clinical files, the causes of brain damage in the organic group were alcohol (26), trauma (8), cerebrovascular accident and multiple sclerosis (2 each), and presenile dementia and brain surgery (1 each). The diagnoses of the controls were alcohol addiction (35); schizophrenia (8); depressive neurosis (7); schizoid personality and habitual excessive drinking (2 each); and drug intoxication, adjustment reaction to adult life, anxiety neurosis, episodic excessive drinking, passive-dependent personality, and inadequate personality (1 each).

Results

Chi-square tests were then run on each MMPI item to determine which items significantly differentiated the two groups. Fifty-six were significant at the .05 level and were, collectively, labeled the *Psychiatric-Organic (P-O) scale*.

The (Form R) items selected, and the directions consistent with an organic diagnosis, were 3, 95, 96, 132, 133, 137, 170, 198, 207, 264, 287, 310, 329, 392, 450, 498, 510, 520 (True); 9, 21, 38, 41, 51, 61, 86, 90, 106, 129, 156, 158, 168, 179, 187, 191, 192, 195, 212, 217, 224, 225, 226, 266, 277, 284, 305, 307, 308, 317, 325, 366, 372, 425, 468, 511, and 540 (False). Eighteen items were keyed "true" and 38 "false."

Scale Characteristics

The internal consistency of the scale was assessed via Ebel's (1951) intraclass correlation technique. It yielded coefficients of .90 for the organics and .68 for the controls. This disparity is puzzling. It may reflect the confusion commonplace in the test-taking performance of psychiatric patients or the heterogeneity of the controls.

A moderate amount of item overlap existed between *P-O* and the other scales. The percentages of the items on the various clinical scales that also appear on *P-O* are detailed in Table 2. Since 10% of all the MMPI items are included on *P-O*, meaningful overlap is represented by the extent to which the percentages exceed 10%. The table reveals that none of the overlaps were high, the largest being only 20%. The greatest overlaps appeared between *P-O* and the Lie (*L*), Psychasthenia (*Pt*), and Schizophrenia (*Sc*) scales. The overlap with Social Introversion (*Si*) was strikingly low.

Correlations were run between the *P-O* scale and each of the MMPI validity and clinical scales for the scale construction samples. These correlations, as well as those of the *P-O* scale with the *Sc-O* scale developed by Watson (1971) to separate organics from schizophrenics, did not differ significantly from group to group. Therefore, the samples were pooled; the correlations for the combined groups are also presented in Table 2. The largest correlations appeared on the validity scales and on *Pt* and *Sc*. Lesser significant correlations appeared between *P-O* and all other MMPI scales studied except Hypochondriasis (*Hs*) and Hysteria (*Hy*). Despite the existence of a slightly higher than chance quantity of overlap between *P-O* and both *Hs*

Table 2
 Percentages of Clinical/Validity Scales Consisting of P-O Items, and Their Correlations with P-O

MMPI scale	Overlapping items		Total	Scale length in items	%	r
	Same direction	Opposite direction				
<i>L</i>	3	0	3	15	20	.55*
<i>F</i>	0	2	2	64	3	-.61*
<i>K</i>	3	1	4	30	13	.62*
<i>Hs</i>	3	1	4	33	12	.08
<i>D</i>	3	5	8	60	13	-.36*
<i>Hy</i>	4	4	8	60	13	-.01
<i>Pd</i>	0	9	9	50	18	-.49*
<i>Mf</i>	1	7	8	60	13	-.52*
<i>Pa</i>	0	4	4	40	10	-.57*
<i>Pt</i>	0	9	9	48	19	-.61*
<i>Sc</i>	2	11	13	78	17	-.57*
<i>Ma</i>	0	6	6	46	13	-.35*
<i>Si</i>	0	1	1	70	1	-.47*
<i>Sc-O</i>	1	6	7	80	9	.54*

* $p < .05$.

and *Hy*, the correlations between *P-O* and those two scales were low. These low correlations may reflect the fact that somatic symptoms are moderately common in both organics and psychiatric patients. The size and breadth of the correlations with those MMPI scales are encouraging, since they suggest that the new scale reflects functional pathology of many sorts. Also encouraging was the presence of a significant correlation between the *P-O* and *Sc-O* scales, both of which were designed to separate organics from functional, although somewhat different, psychiatric samples.

P-O was also moderately correlated with age ($r = .30$); this coefficient can be interpreted as indicating that the scale capitalizes on the tendency for brain-damaged patients to be older than functional patients and makes use of the contribution of age-related personality factors to the discrimination of the two groups from one another.

Cross-Validation

The scale was cross-validated twice with samples drawn from the St. Cloud Veterans Administration Hospital. In the first, the scale was administered to 100 patients, none from the scale construction samples, referred for evaluation of possible brain damage. After testing had been completed, the patient's

physician, nurse, and ward psychologist were consulted. Only those patients in whose cases all three agreed on whether the patient was brain damaged or not were used in the study. *P-O* scores were not available to the raters. Scores of 40 organics and 60 controls were thus collected. As described in their clinical files, the causes of brain damage were alcohol (26 cases); trauma and alcohol (4); trauma, infection, and cerebrovascular accident (2 each); and Wilson's disease, cerebral arteriosclerosis, Sydenham's chorea, and unknown cause (1 each). Preadministered MMPIs were then collected from the subjects' Psychology Service files and compared. The mean *P-O* score of the organics was 31.6 ($SD = 7.7$), whereas that for the controls was 25.4 ($SD = 9.5$). The difference was significant, $t(98) = 3.43$, $p < .005$. This finding was interpreted as being very encouraging, since the samples drawn consisted of subjects in whose case the organic diagnosis was sufficiently problematic to require special evaluation.

As a further test, the scale was scored on a sample of recently admitted/readmitted males under 60 at the St. Cloud Veterans Administration Hospital. According to their clinical file diagnoses, these subjects consisted of 30 organics, 55 neurotics, 98 alcoholics (men with diagnoses of alcohol addition or habitual excessive drinking), 56 character disorders

Table 3

Mean P-O Scores, Standard Deviations, and *t*s for Differences Between Organic and Functional Groups

Sample	<i>M</i>	<i>SD</i>	<i>t</i>
Organic	33.2	6.5	—
Process	26.4	9.3	3.89**
Reactive	27.2	8.2	3.59**
Neurotic	27.2	9.1	3.50**
Alcoholic	27.2	8.8	4.04**
Character disorder	26.8	9.1	3.71**
Affective psychosis	26.8	10.2	2.29*

* $p < .05$.

** $p < .001$.

(men with diagnoses of personality disorder or sexual deviancy), 17 affective psychotics, and 105 schizophrenics. The schizophrenics were split into process and reactive groups by the Ullmann and Giovannoni (1964) Process-Reactive Scale. Those 54 who produced scores of 13 or above were characterized as process, whereas those 51 with scores of 12 or less were classified as reactives. As described in their clinical files, the causes of brain damage in the organic sample were alcohol (10); trauma (8); both trauma and alcohol (3); and Huntington's chorea, atrophy, arteriosclerosis, Wilson's disease, abscess, presenile dementia, cerebrovascular accident, multiple sclerosis, and tumor (1 each). These diagnoses were also made without the benefit of P-O scores. An *F* test run between the means of the seven groups was significant, $F(6, 354) = 2.36$, $p < .05$, and *t* tests were run between the organics and each of the functional groups. These are displayed in Table 3. The reader will note that the mean

for the organics was 33.2, whereas those for the six functional groups ranged from 26.4 to 27.2. All *t*s were significant at at least the .05 level, and five of the six were significant at the .01 level.

P-O Score Interpretation

Although the *t*s are encouraging, the relatively high standard deviations suggest that P-O scale scores between the means of the organic and functional groups are probably of dubious value in the separation of organic from functional patients. Therefore, single cutting scores are not presented here. Instead, distributions of the P-O scores for the 70 organics and 391 functional patients in the two studies were plotted for the purpose of identifying the relative probabilities of organic and psychiatric diagnoses among scorers at various P-O scale levels. These percentages, which are presented in Table 4, allow the diagnostician to make a "better odds" interpretation of P-O scores. For example, given equal base rates, scores between 15 and 20 were 3 times more common among functional than organic patients. Interpretation of these scores, of course, should be tempered by a consideration of local base rate. With even base rates, scores below 27 are more typical of psychiatric than organic patients, whereas those above 27 are more common among brain-damaged patients. Scores between 27 and 32 contribute little to diagnostic differentiation.

Combined Use of P-O and Ability Measures

In at least two studies (Watson, 1978; Watson et al., 1978), the use of personality

Table 4

Percentages of Organic and Functional Subjects at Various P-O Score Ranges and Odds for Organic/Functional Diagnoses

P-O range	Percentages		Odds	
	Organic	Functional	Organic	Functional
39+	25	9	2.78	1
33-38	30	21	1.43	1
27-32	24	21	1.14	1
21-26	15	23	1	1.53
15-20	5	16	1	3.20
0-14	1	10	1	10.00

and ability measures in concert substantially improved on the discriminative power made available by either alone. Following the technique used in those studies, the scores of the 70 organics and 391 controls in the two cross-validation studies on the *P-O* scale and on the Benton Visual Retention Test (BVRT) error scores were graphed on a scatterplot. (The BVRT was used because earlier research [Watson, 1965; Watson, 1968; Watson et al., 1968, 1969] has suggested that it is better able to separate organic from schizophrenic patients than most other visual/motor tests for brain damage.) A straight line was then drawn in such a fashion as to maximize discrimination between the two groups (see Figure 1). This discrimination line yielded hit rates of 69% and 75% for the organic and the control samples, respectively, and an (unweighted) average hit rate of 72% over the two samples. These hit rates are somewhat better than the best that could be obtained using either *P-O* (organics, 77%; controls, 52%; $M = 64.5\%$; cutoff = 27.5) or BVRT (organics, 60%; controls, 77%; $M = 68.5\%$; cutoff = 9.5) scores alone.

Discussion

The results of the cross-validations are generally encouraging. They indicate that the *P-O* scale can be used with moderate accuracy to separate organics from psychiatric patients and that it can improve the prediction available from at least one ability-oriented test. Additionally, the item overlaps with the 13 MMPI validity and clinical scales are low enough to indicate that it makes a contribution to assessment that is not available from any one of those scales.

Additional cross-validations from other laboratories are needed. In particular, the utility of the scale with female samples needs assessment; earlier research (Watson, 1971) indicated that the *Sc-O* scale was of no value in separating organic and schizophrenic females. It appears that the personality correlates of brain damage may vary with sex, and the *P-O* scale's validity as a differentiator of female organic and functional patients should not be assumed.

It is particularly important that readers not

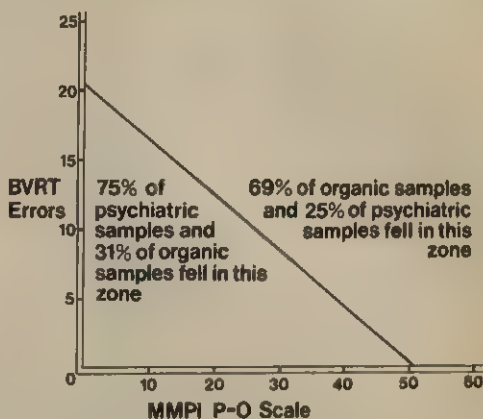


Figure 1. Optimal cutting line for separation of organic and functional patients.

view the *P-O* scale as a technique for identifying brain damage in nonpsychiatric settings. There is no evidence that it will separate organics from psychiatric normals or medical patients, and earlier research has indicated that our *Sc-O* scale does not separate organics from chronic pain or spinal-cord injury patients (Sand, 1973).

Reference Notes

1. Andersen, D. O., Felling, J., & Seitz, F. *The differential diagnosis of schizophrenic and organic patients using the MMPI*. Unpublished manuscript, 1971 (Available from Dennis Andersen, Central Minnesota Counseling Center, St. Cloud, Minnesota).
2. Gilbertstadt, H. *Clinical study and appraisal of some tests related to organicity*. Unpublished manuscript, Veterans Administration Hospital, Minneapolis, Minnesota, 1973.

References

- Ayers, J., Templer, D. I., & Ruff, C. F. The MMPI and the differential diagnosis of organicity vs. schizophrenia: Empirical findings and a somewhat different perspective. *Journal of Clinical Psychology*, 1975, 31, 685-686.
- Brown, E. C., Casey, A., Fisch, R. I., & Neuringer, C. Trail-Making Test as a screening device for the detection of brain damage. *Journal of Consulting Psychology*, 1958, 22, 469-474.
- Ebel, H. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, 16, 407-424.
- Holland, T. R., Lowenfeld, J., & Wadsworth, H. M. MMPI indices in the discrimination of brain-damaged and schizophrenic groups. *Journal of Consulting and Clinical Psychology*, 1975, 43, 426.
- Russell, E. W. Validation of a brain-damage va.

- schizophrenia MMPI key. *Journal of Clinical Psychology*, 1975, 31, 659-661.
- Sand, P. L. Performance of medical patient groups with and without brain damage on the Hovey (O) and Watson (Sc-O) MMPI scales. *Journal of Clinical Psychology*, 1973, 29, 235-237.
- Ullmann, L. P., & Giovannoni, J. M. Development of a self-report measure of the process-reactive continuum. *Journal of Nervous and Mental Disease*, 1964, 138, 38-42.
- Watson, C. G. WAIS profile patterns of hospitalized brain-damaged and schizophrenic patients. *Journal of Clinical Psychology*, 1965, 21, 294-295.
- Watson, C. G. The separation of NP hospital organics from schizophrenics with three visual motor screening tests. *Journal of Clinical Psychology*, 1968, 24, 412-414.
- Watson, C. G. An MMPI scale to separate brain-damaged from schizophrenic men. *Journal of Consulting and Clinical Psychology*, 1971, 36, 121-125.
- Watson, C. G. A simple bivariate screening technique to separate NP hospital organics from other psychiatric groups. *Journal of Clinical Psychology*, 1973, 29, 448-450.
- Watson, C. G., Davis, W. E., & Gasser, B. The separation of organics from depressives with ability- and personality-based tests. *Journal of Clinical Psychology*, 1978, 34, 393-397.
- Watson, C. G., & Klett, W. G. The Henmon-Nelson, Cardall-Miles, Slosson, and Quick Tests as predictors of WAIS IQ. *Journal of Clinical Psychology*, 1975, 31, 310-313.
- Watson, C. G., & Thomas, R. W. MMPI profiles of brain-damaged and schizophrenic patients. *Perceptual and Motor Skills*, 1968, 27, 567-573.
- Watson, C. G., Thomas, R. W., Andersen, D., & Felling, J. Differentiation of organics from schizophrenics at two chronicity levels by use of the Reitan-Halstead Organic Test Battery. *Journal of Consulting and Clinical Psychology*, 1968, 32, 679-684.
- Watson, C. G., Thomas, R. W., Felling, J., & Andersen, D. Differentiation of organics from schizophrenics with the Trail Making, Dynamometer, Critical Flicker Fusion and Light Intensity Matching Tests. *Journal of Clinical Psychology*, 1969, 25, 130-133.

Received October 3, 1977 ■

"Biosyntonic" Therapy: Modification of an Operant Conditioning Approach to Pedophilia

J. Dennis Nolan and Curt Sandman
Ohio State University

A modified operant conditioning approach for treating sexual behavior problems, termed "biosyntonic" therapy, is introduced. Following physiological diagnosis, a child molester was successfully taught to alter his inappropriate physiological arousal to prepubescent females without disrupting his arousal to adult females. In addition to eliminating his child molesting, the procedures also improved the correspondence between his physiological and verbal response patterns and decreased his anxiety. The treatment remained successful at a 6-month follow-up.

Deviant sexual behaviors have been much more resistant to traditional treatments than other behavioral problems (Bieber, 1962). Even aversive conditioning approaches, which seem to be the treatment of choice for some deviant sexual behaviors (O'Leary & Wilson, 1975), are not as effective for sexual behavior problems as for other behavior problems. In fact, radical surgical techniques such as castration have been proposed seriously as the only adequately reliable intervention for such psychopathic sexual behaviors as pedophilia (child molesting).¹ In this study we report the successful application of an approach we have labeled "biosyntonic" to a formerly intractable sexual behavior problem. This approach offers a promising alternative that does not preclude all sexual behavior. It attempts to bridge the hiatus between so-called mental and physical aspects of human behavior generated by traditional approaches to understanding and modifying behavior.

Most forms of psychotherapy rely exclusively on the client's verbal behavior (in most cases the real concern is with some internal

state or "mental" activity) and, except in very rare situations, completely ignore physiological processes. The behaviorally oriented therapist often relies on the observation of behavioral changes in the client, but again, without regard to the physiological state of the organism. Conversely, although scattered reports of the behavioral consequences of conditioning physiological responses exist, for the most part they deal with behaviors of only indirect practical relevance (Beatty, Greenberg, Deibler, & O'Hanlon, 1975; Blanchard & Young, 1973; McCanne & Sandman, 1974). Biofeedback approaches have been partially successful in the treatment of some physical disorders (e.g., functional auricular arrhythmias; Weiss & Engel, 1971), especially when the modified response system is part of the symptom at issue. Attempts to characterize the subtle yet pervasive physiological processes underlying certain behaviors and diagnostic classes have also been reported (Greenfield, Katz, Alexander, & Roessler, 1963; Lacey, 1959; Lacey & Lacey, 1970; McCarron, 1973; Sandman, 1975; Shagass & Schwartz, 1962), but no attempts have involved all of these perspectives

The authors' contributions were complimentary and of equivalent importance. Hence, order of authorship was determined by a coin toss.

Requests for reprints should be sent to either J. Dennis Nolan or Curt Sandman, Department of Psychology, Ohio State University, Columbus, Ohio 43210.

¹ A superior court judge in San Diego, California, recently authorized castration (as an alternative to life imprisonment) of a child molester. The judgment was based on a psychiatrist's report that the molester was aware of what he was doing and could not be cured of his perversion.

in the treatment of serious human behavior problems.

Rationale for Modification

In most aversive conditioning approaches to modifying undesirable behavior, the conditioning process involves pairing an aversive stimulus with the target stimulus (e.g., O'Leary & Wilson, 1975). For example, in the case of a pedophile, electric shock might be systematically paired with slides of children or perhaps with the patient's own imagery involving a child. In more sophisticated uses of such procedures, delivery of the aversive stimulus may be on a variable rather than a fixed schedule (e.g., Marks & Gelder, 1967), but the delivery of the aversive stimulus is invariably contingent on the presentation of the target stimulus, not on the patient's physiological response to that stimulus. Hence, in the usual aversive conditioning paradigm, shock is delivered regardless of whether (or when) the patient's response to the target stimulus is inappropriate. In the approach described in this study, the reinforcing stimulus is delivered only when the patient's physiological response to a target stimulus suggests that he is aroused. This approach insures that conditioning is problem specific yet does not depend directly on the patient's verbal report of his arousal. The reinforcing stimulus can be delivered immediately and accurately at the very onset of the physiological response without relying on either the accuracy or the speed of the patient's verbal report. Even if the verbal report were considered accurate, it probably would not occur as quickly or as reliably as the physiological response. Any resulting delay or unreliability of the delivery of the reinforcing stimulus could weaken the effectiveness of the conditioning procedure. The "biosyntonic" approach avoids both of these potential problems. An additional advantage of this approach is that the reinforcing stimulus can be initiated or terminated *during* the presentation of a target stimulus whenever changes in the patient's physiological response warrant such action.

We have introduced the term "biosyntonic" to refer jointly to the synchrony among physiological systems and to the correspondence

between thoughts or attitudes and the physical state of the organism. That is, we are concerned with the syntony among verbal (mental), physiological, and gross behavioral aspects of the human experience. A major assumption of this approach is that attitudes, emotions, and behavior are intimately related to the physiological state of the organism. The physiology is not considered a mere passive reflection of the psychological state but rather an active (though probably incomplete) determinant of attitudes, emotions, motives, and beliefs. Thus, it is our assumption that behaviors and all their attendant psychological constructs are represented by distinct physiological patterns, each pattern relating to a different psychological state. To further complicate the matter, a sizable body of research (Hein, 1969; Lacey & Lacey, 1958; McCanne & Sandman, 1975; Sandman, 1975) suggests that such patterns are highly idiosyncratic to both persons and situations. From such a "biosyntonic" perspective, then, the work of the therapist is to carefully identify (diagnose) the physiological state of the individual in specific situations before attempting to intervene. Once the relationship between mental and physical activity has been observed in nonproblem areas, then the therapist can approach the problem areas searching for evidence of disharmony in this relationship. The goal is to target physiological systems in the problem area that appear to be disparate with the patterns of responses in the more normal areas of the client's life. Once a physiological system has been observed to be abnormal in the sense described herein, the "biosyntonic" position implies that by changing the physiological pattern, the client can be helped to experience a change in attitude or emotions or both and, concomitantly, to experience a change in overt behavior. Although several studies conducted in our laboratory with normal subjects support this view (Baker, Sandman, & Pepinsky, 1975; Kaiser & Sandman, 1975; McCanne & Sandman, 1974), the present case study presents the first successful clinical test of our thesis.

Case Report

Mr. J., a 32-year-old blue-collar worker, had a history of numerous sexual experiences

with children (nearly all females) dating from his teens. Fearful of the personal and legal consequences of the discovery of his experiences with a female child, he consented to the procedures described below as a last resort. Traditional approaches had been tried without success.

Assessment

Preliminary diagnostic information indicated extreme sexual attraction for prepubescent females, although Mr. J. maintained an adequate sexual relationship with his wife. An extensive pretreatment physiological assessment was conducted by attaching electrodes for the measurement of heart rate, peripheral vasomotor activity, respiration, and skin potential. Mr. J. was reclined in a comfortable chair and viewed a standard sequence (Sandman, 1975) of pictures presented on a screen in front of him. The series included sexually arousing, neutral, and highly distressing (e.g., mutilated corpses) slides. He rated each of them on a 9-point pleasure-stress scale. Physiological recording was done with a Grass Model 7B polygraph equipped with appropriate preamplifiers and driver amplifiers that were housed in the adjacent control room.

Mr. J. exhibited differential physiological responses to pleasurable, neutral, and unpleasurable stimuli only in heart rate. His vasomotor and electrodermal responses were not differentially responsive to stimuli of different pleasure ratings. Hence, we hypothesized that the cardiovascular response system could be a significant component of his total response to pleasurable stimuli. We therefore focused on heart rate in a discriminative conditioning paradigm used throughout 16 sessions; the conditioning was designed to alter his physiological responses to female children without disrupting his responses to adult females.

Initially, four sets (male and female adults and male and female children) of six stimuli each were chosen. Mr. J. was asked to bring in pictures of preadolescent girls that he found arousing to varying degrees. Very prevalent among the most arousing stimuli were pictures taken from mail-order catalogues of girls modeling underwear. Upon request he also brought

comparable pictures of preadolescent boys, though he claimed none of these were highly arousing. He selected adult male and female stimuli from a group of slides available at the treatment center. Mr. J. rated each of the stimuli in each of the four sets in terms of his perceived sexual arousal to them. Ratings were based on the 9-point pleasure-stress scale described earlier.

The most striking feature of his physiological response pattern to these four sets of stimuli was the extremely elevated heart rate response to the pictures of the semiclothed female children. However, unlike his response to nude adult females, his verbal report of arousal to the slides of children was not consistent with his physiological response to them² (Figure 1A). Neither his verbal nor his physiological responses suggested arousal to male children or male adults. Therapy was therefore designed to accomplish both the reduction of the inappropriate arousal to female children and improvement of the congruity between the patient's verbal and physiological responses to such children.

Results of Treatments

In Phase 1 of the treatment, heart rate increases during the presentation of slides of young female children were punished by the administration of electric shock (4–6 mA) to Mr. J.'s index finger. Heart rate was detected by level detectors (BRS Digibit Logic Modules) with preset criteria. The detectors were connected to a stimulator (Chicago-Nuclear) that automatically delivered electric shock when Mr. J.'s heart rate exceeded the criterion. The criterion for heart rate was the 90th percentile of resting heart rate range. Periodic adjustments were made for basal heart rate shifts. As illustrated in Figure 2a, throughout the first session there was a substantial difference between the heart rate dur-

² An analytic interpretation of this finding could be that the patient is employing a psychological defense, denial or repression, and thereby is not "able" to consciously report his true feeling about the stimuli. Lazarus (1968) has demonstrated that the process of denial can even effect the physiological responses to stimuli.

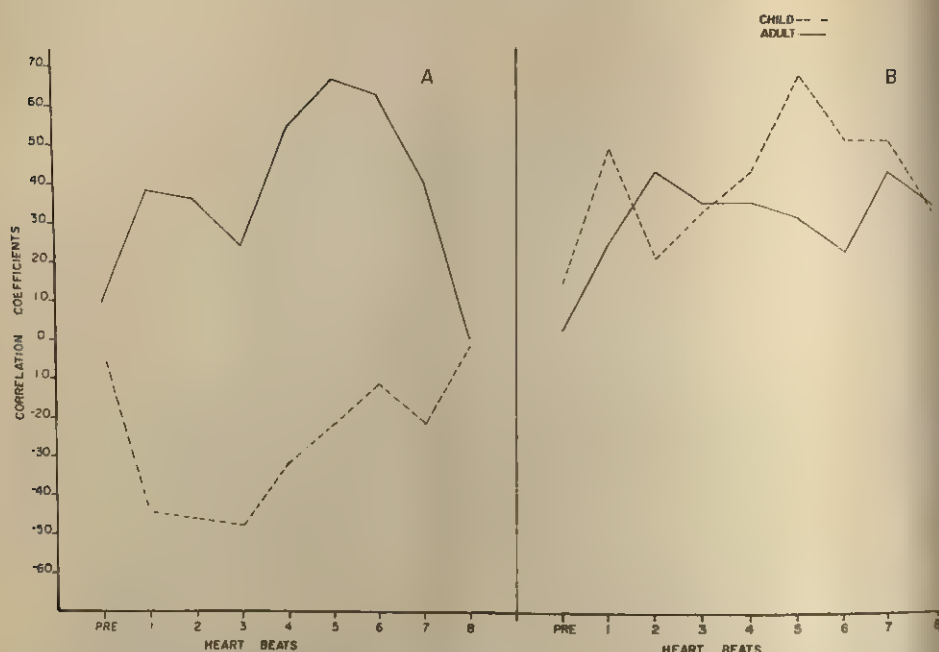


Figure 1. Correlation coefficients between verbal ratings of arousal and heart rate during presentation of slides and mature and immature females (1A) before and (1B) after "biosyntonic" therapy. (Data are for the first eight heartbeats of each stimulus presentation.)

ing presentation of slides of female children and the rate during slides of adult females ($p < .05$ for beats 1 and 6; $p < .01$ for beats 7 and 8).⁸ By the second session, however, the abnormal response to the children had been eliminated (Figure 2b; no significant differences). Responses to the two male stimulus classes remained unchanged, and we therefore discontinued presentation of slides of males at this point. To guard against the possible effects of habituation, during the third session new slides of semiclothed female children were introduced without compromising the effects of therapy. At the end of this session, Mr. J. reported the first incident in which he experienced a sharp pain in his index finger (where the shock had been delivered) whenever he felt the beginning of what he termed his "automatic arousal" response to observing preadolescent girls in provocative positions during his daily activities. The effect of this conditioned pain response was the abortion of the automatic arousal response and the complete elimination of the overt child molesting that had previously occurred under

such conditions. He also reported that he could no longer find pictures of young girls that were sexually arousing in any of the catalogues or magazines he had used previously. This report was considered a very positive sign, but continued treatment was nevertheless considered essential. Therefore, we introduced a new and potentially more provocative set of slides of a female child in poses ranging from fully clothed to nude. Mr. J.'s dramatic heart rate response to the very provocative stimuli is illustrated in Figure 2c. Our Phase 1 procedures were again effective in reducing Mr. J.'s responses to those powerful stimuli (Figure 2d), but they did not eliminate the responses ($p < .05$ for beats 6 and 7; $p < .01$ for beats 1, 2, 3, 4, 5, and 8). A more powerful conditioning technique was needed.

In Phase 2 we attempted to enhance the

⁸ One-tailed t tests were performed on beat-by-beat differences in heart rate during slides of adult females versus slides of young girls. The p values are listed for all statistically significant results.

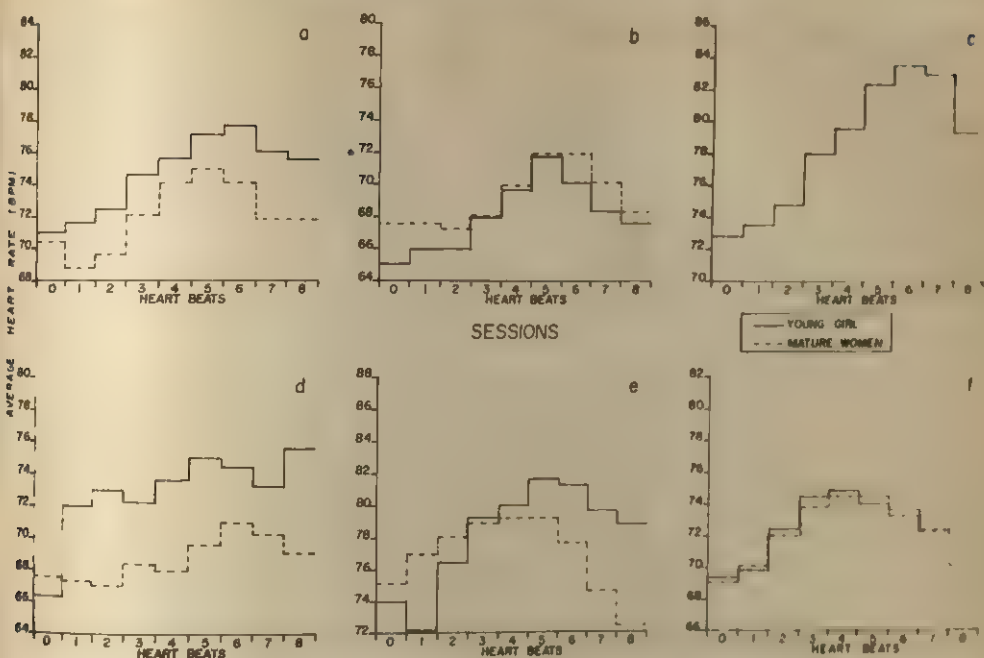


Figure 2. Heart rate patterns during presentation of provocative pictures of mature and immature females through the course of treatment. (Data are for the first eight heartbeats of each stimulus presentation.)

discrimination by providing positive reinforcement (monetary reward based upon the amount of time Mr. J. kept his heart rate above criterion) for appropriate heart rate responses (i.e., responses that were consis-

tent with the responses of the diagnostic sessions) to the adult females while continuing punishment for heart rate increases to pictures of the child. In Figure 2e it is evident that the pervasive effect of the provocative

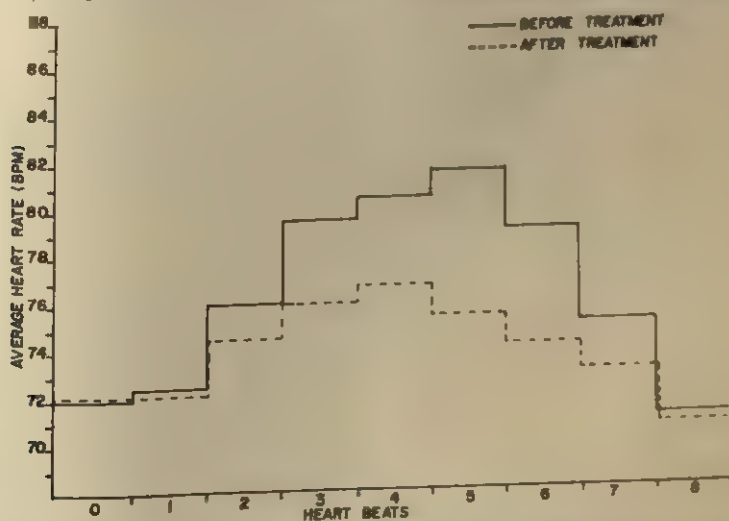


Figure 3. Heart rate response to prepubescent females before and after "biosyntonic" therapy. (Data are for the first eight heartbeats of each stimulus presentation.)

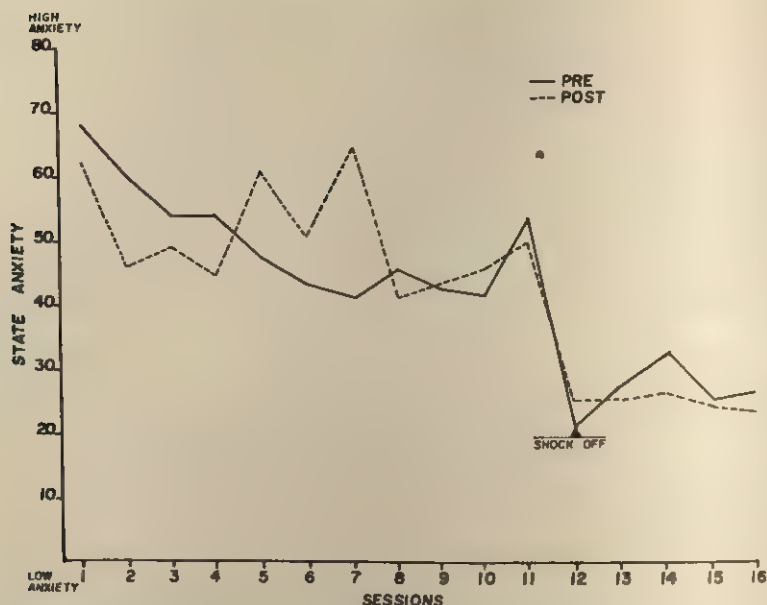


Figure 4. State anxiety scores for each session before and after "biosyntonic" therapy.

child had diminished and the acceleratory response had changed by 9 beats. However, even though this effect was dramatic and substantial, we were still somewhat dissatisfied with the discrimination ($p < .05$ for beat 6; $p < .01$ for beats 7, 8). It seemed as if the aversive conditioning procedures, though initially essential in establishing the reinforcement contingencies, also might have contaminated both the heart rate response and the discrimination. The usual response to shock, for example, is heart rate acceleration. In order to eliminate this possible contamination, we introduced Phase 3.

In Phase 3, positive reinforcement remained contingent upon heart rate acceleration to adult females, but positive reinforcement was also presented for *inhibition* of the acceleratory heart rate response to slides of the young girl. This procedure proved highly effective within two sessions, as illustrated in Figure 2f (no significant differences). We continued the Phase 3 procedure for two additional sessions with no incidents and terminated treatment after arranging for follow-up evaluations.

At the end of the therapy the patient's extreme heart rate accelerations to pictures of young female children had been eliminated

(Figure 3),⁴ and the correspondence between his verbal and physiological response patterns was normalized (Figure 1B). He had not assaulted a child for several months (compared with a history of almost weekly incidents).

In addition to the physiological data and the patient's verbal reports, state anxiety (Spielberger, Gorsuch, & Lushene, 1969) data were collected for each session, both before and after treatment. There was a consistent decrease in anxiety over the 16 sessions,⁵ with a marked drop at Session 12, apparently reflecting the elimination of the shock contingency (Figure 4).

Follow-Up

Six months after treatment was terminated, Mr. J. reported that he had not experienced

⁴To equate starting heart rate, 7 beats/minute were subtracted from the rate for each "before" beat. The t tests on the adjusted comparisons were statistically significant for beats 3, 4 ($p < .05$), and 5 ($p < .01$).

⁵Using $SD = 14$, based on norms given by Spielberger, Gorsuch, and Lushene (1970), this decrease was statistically significant, $t(15) = 2.78$, $p < .01$.

any incidents of child molesting. He remarked that when he was in a situation formerly considered provocative, he still experienced some pain in his index finger. Of more importance, he has consciously avoided compromising situations, indicating that he is able to control his behavior much more efficiently than before "biosyntonic" therapy. It was of special interest to learn that Mr. J. attends X-rated movies at a far higher rate than he did before therapy. He had been to only one such movie before therapy, but in the first 6 months following therapy, he attended about nine. Whether this is the direct consequence of positively reinforcing his physiological responses to nude adult women in our procedure, or a compatible symptom substitution, is uncertain. In any case, the result is that a socially condoned activity has apparently replaced his psychopathic and deviant behavior.

Discussion

The new procedures reported here suggest a dramatically successful alternative to the treatment of an intractable psychopathic sexual behavior problem. The important procedural distinction between the "biosyntonic" therapy approach and other conditioning approaches is that in "biosyntonic" therapy the delivery of reinforcement is contingent on the appropriateness of the patient's physiological response to a stimulus and not merely on the presentation of the stimulus. In the present case, "biosyntonic" therapy insured that the conditioning was problem specific—even though it was not directly dependent on the patient's verbal report of his arousal. Apparently, it affected the underlying state in such a way that only the psychopathic component of his sexual drive was altered.

We are convinced that the success reported here is largely dependent on the careful diagnostic evaluations that allowed the identification of a specific physiological system (in this case, heart rate) as central to the psychopathic behavior problem in terms of both its differential responsiveness to stimuli and the disparity between the physiological and verbal indices of arousal. If correspondingly careful diagnostic work indicates comparable physio-

logical substrates to other behavioral problems, a "biosyntonic" approach to therapy may offer a viable treatment approach to other problems.

References

- Baker, W. M., Sandman, C. A., & Pepinsky, H. B. Affectivity of task, rehearsal time and physiological response. *Journal of Abnormal Psychology*, 1975, 84, 539-544.
- Beatty, J., Greenberg, A., Deibler, W. F., & O'Hanlon, J. F. Operant control of occipital theta rhythm affects performance in a radar monitoring task. *Science*, 1975, 183, 871-873.
- Bieber, J. *Homosexuality*. New York: Basic Books, 1962.
- Blanchard, E. B., & Young, L. O. Self-control of cardiac functioning: A promise as yet unfulfilled. *Psychological Bulletin*, 1973, 79, 143-162.
- Greenfield, N., Katz, D., Alexander, A., & Roessler, R. The relationship between physiology and psychological responsivity: Depression and galvanic skin response. *Journal of Nervous and Mental Disease*, 1963, 136, 535-539.
- Hein, P. L. Heart rate conditioning in the cat and its relationship to other physiological responses. *Psychophysiology*, 1969, 5, 455-464.
- Kaiser, D. N., & Sandman, C. A. Physiological patterns accompanying complex problem solving during warning and nonwarning conditions. *Journal of Comparative and Physiological Psychology*, 1975, 89, 357-363.
- Lacey, J. I. Psychophysiological approaches to the evaluation of psychotherapeutic process and outcome. In E. A. Rubenstein & M. B. Parloff (Eds.), *Research in psychotherapy*. Washington, D.C.: American Psychological Association, 1959.
- Lacey, J. I., & Lacey, B. C. Verification and extension of the principle of autonomic response stereotypy. *American Journal of Psychology*, 1958, 71, 50-73.
- Lacey, J. I., & Lacey, B. C. *Physiological correlates of emotion*. New York: Academic Press, 1970.
- Lazarus, R. S. Emotions and adaptation: Conceptual and empirical relations. In W. J. Arnold (Ed.), *Nebraska Symposium on Motivation* (Vol. 16). Lincoln: University of Nebraska Press, 1968.
- Marks, I. M., & Gelder, M. G. Transvestism and fetishism: Clinical and psychological changes during faradic aversion. *British Journal of Psychiatry*, 1967, 113, 711-729.
- McCanne, T. R., & Sandman, C. A. Instrumental heart-rate responses and visual perception: A preliminary study. *Psychophysiology*, 1974, 11, 283-287.
- McCanne, T. R., & Sandman, C. A. Determinants of human operant heart-rate conditioning: A systematic investigation of several methodological issues. *Journal of Comparative and Physiological Psychology*, 1975, 88, 609-618.

- McCarron, L. T. Psychophysiological discriminants of reactive depression. *Psychophysiology*, 1973, 10, 223-230.
- O'Leary, K. D., & Wilson, G. T. *Behavior therapy: Application and outcome*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.
- Sandman, C. A. Physiological responses during escape and nonescape from stress in field independent and field dependent subjects. *Biological Psychology*, 1975, 2, 205-216.
- Shagass, C., & Schwartz, M. Cerebral cortical reactivity in psychotic depressions. *Archives of General Psychiatry*, 1962, 6, 235-242.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *The State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1969.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *Manual for the State-Trait Anxiety Inventory (Self Evaluation Questionnaire)*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Weiss, T., & Engel, B. Operant conditioning of heart rate in patients with premature ventricular contractions. *Psychosomatic Medicine*, 1971, 33, 301-321.

Received October 5, 1977 ■

Brief Reports

Female Criminal Violence and Differential MMPI Characteristics

Patricia B. Sutker, Albert N. Allain, and Scott Geyer
Department of Psychiatry and Behavioral Sciences
Medical University of South Carolina

A cross-validated approach was used to compare Minnesota Multiphasic Personality Inventory (MMPI) scale elevations and profile patterns produced by female murderers and nonviolent offenders in two geographic regions. Murderers from both prison sources produced subdued group mean profiles, whereas nonviolent offenders were characterized by elevations on Scale 4. Discriminant function classification was highly dependent on scores on Scales 4, 5, K, and A and correctly identified 82% of violent and 78% of nonviolent offenders. A principal-components analysis yielded five components or profile types, but only the component defined by high positive loadings for Scale 4 differentiated between the groups.

Representing perhaps the antithesis of traditional womanhood, female aggressivity would seem to hold great interest for investigators. Yet few studies have described psychosocial factors that differentially characterize violent and nonviolent women (Climent, Rollins, Ervin, & Plutchik, 1973; Cole, Fisher, & Cole, 1968). More often, empirical studies of the psychological parameters of violent behavior among men have been undertaken, with the work of Megargee and associates as the most comprehensive example (Megargee, Cook, & Mendelsohn, 1967; Megargee & Hokanson, 1970). Assuming that certain personality or cognitive characteristics may be predictive of female criminal violence, the present study used a cross-validated approach to identify personality dimensions or patterns measured by the Minnesota Multiphasic Personality Inventory (MMPI) that might be related to extreme violent behavior among female felons.

Subjects were women convicted for murder and nonviolent offenses in the Louisiana Correctional Institute for Women (LCIW) and the Women's Correctional Center (WCC) in South Carolina. Selection in both prisons involved the following steps: random identification of roughly one third of the names from prison rolls (150 in LCIW; 180 in WCC), systematic reduction and

replacement of women unable to read sufficiently well to complete the instruments (10%-15% of those initially selected) or unwilling to participate (2%-5%), and posttest exclusion of women convicted for violent crimes less extreme than murder or manslaughter for the violent group and of women convicted for nonviolent offenses with a record of violent crimes for the nonviolent group. Both samples also excluded newly inducted inmates (imprisoned less than 5 months). The final breakdown of LCIW women ($n = 32$) included 12 violent and 20 nonviolent offenders, and of the 30 women selected at WCC, 10 had been convicted of murder or manslaughter and 20 of drug or property offenses. Distribution by race was equal for violent and nonviolent categories in both prison samples, and groups did not differ on personal or history variables including age, education, age at current offense, months served on current offense, total time incarcerated, and intellectual level. Data collection instruments were the Shipley Institute of Living Scale, Raven's Progressive Matrices, the MMPI, and a structured private interview. Sixteen MMPI scales were scored and converted to T -score values: 10 clinical, 3 validity, Welsh's A and R , and Barron's Ego Strength scales. The revised Overcontrolled Hostility (O-H) scale (Megargee et al., 1967) was also scored, but no T -score transformation was applied.

Violent criminal offenders responded to the MMPI in a less deviant fashion than nonviolent felons, and group mean profiles fell completely within the "normal" range. LCIW group comparisons showed significant differences on Scale 4

Requests for reprints and for an extended report of this study should be sent to Patricia B. Sutker, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina, 171 Ashley Avenue, Charleston, South Carolina 29403.

alone, $F(1, 30) = 4.29$, $p < .05$, and differences for WCC were found on Scales 4, $F(1, 28) = 6.10$, $p < .05$, and 2, $F(1, 28) = 5.29$, $p < .05$. Between-group comparisons for prison groups revealed only one difference—WCC murderers produced more elevated scores on Scale L, $F(1, 20) = 6.76$, $p < .05$, than LCIW murderers. The two prison samples were combined for subsequent analyses. For combined samples, murderers scored lower on Scale F, $F(1, 60) = 4.45$, $p < .05$, and Scale 4, $F(1, 60) = 10.09$, $p < .01$, and higher on Scale K, $F(1, 60) = 6.07$, $p < .05$, and Scale 5, $F(1, 60) = 5.09$, $p < .05$. Comparisons on the O-H scale showed no significant differences between violent and nonviolent groups, $F(1, 60) = .95$, $p > .05$. Prediction of group identity using stepwise discriminant function analysis was highly accurate (82% of violent and 78% of nonviolent offenders were classified), with categorization heavily dependent on scores on Scales 4, 5, K, and A.

Principal-components analysis with varimax rotation identified five independent components with eigenvalues greater than 1.00, which accounted for 78% of the total variance. Components represented anxiety maladjustment, activity, conversion reaction, masculinity-femininity, and social deviance dimensions. Group comparisons of mean factor scores showed that only the factor suggesting social deviance, defined by high positive loadings on Scale 4 and moderate negative loadings on Scale L, was differentially related to offender classification, $F(1, 60) = 9.44$, $p < .01$. Using the Meehl system of profile classification, 50% of the murderer profiles were categorized as normal, as compared to 25% of nonviolent offenders. Within the nonviolent offender group, 42.5% of profiles were classified as conduct disorder, 30% as psychotic, and 2.5% as neurotic, whereas only 14% of murderer profiles were labeled *conduct disorder*; 36%, *psychotic*; and 0%, *neurotic*. The most prevalent code type among murderers was 4-5/5-4 (23%), with 1-6/6-1, 2-8/8-2, and 6-8/8-6 each accounting for 10% of the profiles. In contrast, the 4-5/5-4 type was characteristic of 5% of the nonviolent offenders. There was only one 4-9/9-4 profile among murderers, but this type accounted for 30% of nonviolent offender profiles.

Results suggest that there are significant and reliable relationships between extreme criminal violence and aspects of MMPI performance. Women convicted of criminal homicide tended to respond to MMPI items in a manner reflective of minimal involvement in a socially deviant life-style, with reluctance to admit unusual psy-

chological symptoms, limited personal insight, moderate levels of interpersonal anger, and less identification with culturally defined femininity than nonviolent offenders. Thus, women who murdered could be said to be more defensive, less in touch with impulses to action, more socially conforming, and more removed from a stereotyped definition of femininity.

These data also suggest a negative relationship between membership in a violent offender group and sociopathic psychopathology—a finding compatible with the notion that so-called aggressive psychopaths are found infrequently to be extremely assaultive and rarely to kill (Gibbens, Pond, & Stafford-Clarke, 1959). Findings also describe a personality constellation in agreement with the Megargee et al. (1967) hypothesis that overcontrolled persons may more often commit more violent crimes than those who are undercontrolled, although no differences were observed on the O-H scale developed by Megargee et al. using male samples to measure overcontrolled hostility.

That the masculinity-femininity dimension may be implicated as associated with female criminal violence is interesting, in that homicide is a crime most easily identified with masculine aggressive behaviors. Though Scale 5 may be a poor descriptor of social posture regarded as feminine, masculine, or both, its significance in this study underscores the need for further research focusing on female attitudes toward self, men, and society and the relationship of such attitudes toward assertive behavior. Finally, results point to a striking similarity in MMPI profile configurations across samples, with particular focus on differences on Scale 4, a preponderance of normal profiles among murderers, and concentration of conduct disorder classifications for nonviolent offenders. These findings neither generalize to less deviant populations nor predict violent behavior in general. However, they target areas for future research and suggest the need for studies to identify the various personality or cognitive factors potentially predictive of inappropriately aggressive behavior among women in less deviant groups. Such investigations would also have direct implications for the myriad of female assertiveness training classes springing up across the nation.

References

- Climent, C. E., Rollins, A., Ervin, F. R., & Plutchik, R. Epidemiological studies of women prisoners. I. Medical and psychiatric variables related to violent

- behavior. *American Journal of Psychiatry*, 1973, 130, 985-990.
- Cole, K. E., Fisher, G., & Cole, S. S. Women who kill: A sociopsychological study. *Archives of General Psychiatry*, 1968, 19, 1-8.
- Gibbens, T. C. N., Pond, D. A., & Stafford-Clarke, D. A follow-up study of criminal psychopaths. *Journal of Mental Science*, 1959, 105, 108-115.
- Megargee, E. I., Cook, P. E., & Mendelsohn, G. A. Development and validation of an MMPI scale of assaultiveness in overcontrolled individuals. *Journal of Abnormal Psychology*, 1967, 72, 519-528.
- Megargee, E. I., & Hokanson, J. E. *The dynamics of aggression*. New York: Harper & Row, 1970.

Received December 16, 1976 ■

Turning On or Turning Off: Sensation Seeking or Tension Reduction as Motivational Determinants of Alcohol Use

Raymond M. Schwarz, Barry R. Burkhart, and Samuel B. Green
Auburn University

To assess the relative influences of sensation-seeking or tension-reduction motives on drinking behavior, 242 college students completed the Sensation Seeking Scale, the S-R Inventory of General Trait Anxiousness, and a self-report index of drinking behavior. Using correlational and multiple regression procedures, the data consistently indicated a strong positive relationship between sensation seeking and alcohol use, whereas the relationship between anxiety and alcohol use was nonsignificant. The importance of sensation-seeking motives to a comprehensive motivational theory of alcohol is discussed.

Physiologically, alcohol functions to depress central nervous system activity. This decrease in cortical arousal as a primary consequence of alcohol consumption has served as a starting point for much of the psychologically oriented theory about drinking behavior, because it has been assumed that the psychological motivation for alcohol use must complement or reflect these physiological effects. The concept of tension reduction seemed to be the best psychological analogue to the decrease in cortical arousal occurring as a function of drinking; that is, drinking leads to decreased cortical arousal, which results in a decrease in tension anxiety.

Recent research addressed to the alleged relationship between anxiety and alcohol use has returned mixed findings. Alcoholic populations generally have higher levels of self-reported anxiety than normals; however, other studies have found no relationship between self-reported anxiety and alcohol use with college students. To further complicate matters, several studies have demonstrated that consumption of alcohol is associated with increased mood disturbances.

In an attempt to explain these discrepant findings, it has been suggested that researchers distinguish between the tension-reducing effects of alcohol and the fact that organisms may or may

not drink alcohol for its tension-reducing effects. It may be that even if alcohol is a pharmacological sedative, people drink for other reasons.

One such reason can be derived from the concept of optimal level of stimulation. Given that alcohol use is not clearly or consistently related to tension-reduction needs, it may be that a comprehensive theory of alcohol use requires attention to stimulus-seeking needs. Studies of premorbid personality characteristics of problem drinkers have shown many of them to be impulsive, restless, and nonconforming; this description bears a marked resemblance to the characteristics of high stimulus seekers. Additionally, some studies have found significant correlations between drinking behavior and measures of stimulus seeking.

The present study was designed to examine the relationship between tension reduction and sensation-seeking needs and alcohol use. If drinking behavior is a consequence of one or both of these motives, then the amount and frequency of alcohol use should be correlated with measures of the motivational predisposition.

The subjects for this study were 242 undergraduates, both male ($n = 130$) and female ($n = 112$). Each student was asked to complete the Sensation Seeking Scale (SSS); the S-R Inventory of General Trait Anxiousness (S-R GTA); and an index of the individual's drinking behavior, which was based on both frequency and amount of alcohol use. The SSS is composed of five interpretable factor scales: Thrill and Adventure Seeking (TAS), Disinhibition (Dis), Experience Seeking (ES), Boredom Susceptibility (BS), and a General scale of sensation seeking.

This article is based on a thesis submitted by the first author and directed by the second author in partial fulfillment of the requirements of the master's degree from Auburn University.

Requests for reprints should be sent to Barry R. Burkhart, Department of Psychology, Auburn University, Auburn, Alabama 36830.

The S-R GTA is a self-report inventory designed to measure anxiety across four general stimulus situations: interpersonal, physical danger, ambiguous, and routine situations.

The zero-order correlations between the SSS and the drinking index are presented in Table 1. Significant positive correlations were obtained between all five scales and the drinking index. The zero-order correlations between the drinking index and anxiety scales were nonsignificant, with the single exception of the correlation between the Physical Danger anxiety scale and the drinking index. However, this relationship was weak ($r = -.15$, $p < .05$) and, more noteworthy, in the negative direction.

To determine the relative contributions of sensation seeking and anxiety-reduction motives to drinking behavior, several multiple regression analyses were computed. The drinking index was predicted significantly by the SSS, $F(5, 236) = 23.68$, $p < .001$; by the S-R GTA, $F(4, 237) = 2.64$, $p < .05$; and by both combined, $F(9, 232) = 13.58$, $p < .001$. In examining the relative contribution of both scales, it was found that when the SSS scores were stepped into the regression equation first, they accounted for 33% of the total variance of the drinking index scores. The anxiety scales contributed only an additional 1% to the total variance accounted for by the SSS scales. When the anxiety scores were entered into the regression analysis first, they accounted for only 4% of the total variance in predicting the drinking index scores. The addition of the SSS raised the variance accounted for to 34%. In both analyses, the Disinhibition scale added more to the total R^2 than the other scales combined.

The results of the present study can be easily summarized. Alcohol use was not correlated positively with the level of self-reported anxiety even when anxiety was considered as a multidimensional construct. On the other hand, alcohol consumption was strongly related to stimulus-seeking needs, particularly needs to engage in disinhibited forms of sensation seeking. These results, in conjunction with other recent findings, do not support a strict tension-reduction hypothesis and suggest that a new analysis of the motivational determinants of alcohol use is in order.

It should be noted that these data are specific to the young adult population represented by our sample. Moreover, studies with older, alcoholic populations often offer support for an association between anxiety and alcohol use. In combination, these results suggest that there may be a partial developmental course in the relation-

Table 1

Correlations Between Sensation Seeking Scales and Drinking Index Scores

Scale	Drinking index
General	.16**
Thrill and Adventure Seeking	.12*
Experience Seeking	.32***
Disinhibition	.59***
Boredom Susceptibility	.27***

Note. $N = 242$.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

ship between sensation seeking, anxiety, and alcohol use. For young adults, drinking may serve primarily as an outlet for sensation-seeking needs, whereas, later in life, especially for vulnerable individuals, drinking may come to serve as a coping mechanism for feelings of anxiety and stress.

Among the various dimensions of sensation seeking, the most powerful predictor of drinking behavior was the Disinhibition scale, which is best described as measuring a need for extraverted, hedonistic social involvement. The question that obviously presents itself is, "Why does a pharmacological sedative serve as a releasing mechanism for socially extraverted, sensation-seeking behaviors?"

Our hypothesis is that alcohol disinhibits behavior, because of the well-established cultural expectancy that drinking leads inevitably to an inability to exercise moral or social restraint, not just because of its pharmacological effects. Recent research, which found that alcohol disinhibits behavior only if the subjects were told that they were drinking alcohol, offers some support for this hypothesis. Drinking, in effect, provides a culturally sanctioned "time-out" from social control, during which exhibitionistic, hedonistic behavior may be expressed with impunity. It follows, therefore, that individuals with a strong need to engage in disinhibited behavior will be more likely to turn to drinking because of this culturally established expectancy.

In summary, alcohol may be described as serving as a "releaser" for normally restrained social behaviors. If so, then theories of alcohol use will need to take into account the reinforcement potential provided by access to such disinhibited states, especially in individuals with strong motivational needs for sensation seeking.

Received April 18, 1977 ■

Therapist and Client Perceptions of Alternative Roles for the Facilitative Conditions

Robert B. Slaney
University of Akron

The present study examined therapist and client perceptions of one of two transcripts of psychotherapy—one using the facilitative conditions as a treatment and one using them as intermediate variables leading to the suggestion of assertive training. Data were obtained from 100 therapists-in-training and 50 clients. For the therapist group, assertive training was estimated as more effective and the behavioral therapist was seen as more expert and appealing. No differences were found on therapist understanding. For the client group, no significant differences were found. The findings are discussed with their implications for future research.

The presence of the facilitative conditions (Carkhuff, 1969) in current research, writing, and professional and paraprofessional training programs supports Bergin's statement that "their presence and influence is ubiquitous" (Bergin & Suinn, 1975, p. 521). However, while some theorists, such as Carkhuff, seem to place emphasis on the conditions as a primary mode of treatment, others (e.g., Lazarus, 1971) stress the intermediate relationship-enhancing aspects of these variables.

The present study used two groups, therapists and clients, to examine perceptions of the role of the facilitative conditions in psychotherapy. For the therapist group the participants were 50 male and 50 female graduate students in rehabilitation counseling, counseling psychology, and clinical psychology who were involved in training programs in mental health facilities in the Syracuse, New York, area. Data were gathered over a 2-year period. The distribution of students was rehabilitation, 38; counseling, 23; and clinical, 39. All participants had had experience in the treatment of clients. They were randomly assigned to treatment conditions by sex and area of specialization. The mean age was 26.36 years.

Preparation of this article was partially supported by Veterans Administration research Grant 7979-01 while the author was a staff member at the Syracuse Veteran's Administration Hospital.

The author wishes to thank Harvey A. Taub for his helpful comments on an earlier draft of this manuscript.

Requests for reprints and for an extended report of this study should be sent to Robert B. Slaney, Department of Psychology, University of Akron, Akron, Ohio 44325.

The mean educational level was 17.53 years. There were no significant differences based on age or years of education.

A written transcript (Slaney, 1977) represented an excerpt from a therapy session. The transcript consisted of nine client-therapist interactions. The client statements portrayed a male who was having difficulties at work, at home, and in social situations because he was anxious, unsure of himself, and unassertive. The therapist responses were designed to represent facilitative responses using Carkhuff's (1969) Empathy Scale. Two independent raters, experienced in the use of Carkhuff's scale, rated all but one response as meeting or exceeding the criterion of 3.0. A Pearson product-moment correlation of the ratings was .86.

A second transcript was devised, which differed from the first only in the last therapist response. Instead of a facilitative response, the suggestion of assertive training was made. This response was rated below 3.0 on Carkhuff's (1969) scale. A group of seven raters, all experienced therapists with PhDs, were asked to judge the appropriateness of the use of assertive training with this client. A 7-point rating scale ranging from 1 (very inappropriate) to 7 (very appropriate) was used. The mean rating was 5.6.

The transcripts were randomly distributed. After reading them, subjects completed a rating scale, which included an estimate of the effectiveness of the eventual outcome of the treatment and three therapist characteristics: expertness, understanding, and appeal. The scales were 8-point Likert-type scales ranging, for example, from extremely inexpert to extremely expert.

The mean ratings were slightly to moderately positive. None was negative. A 2×2 (Treatment \times Sex) analysis of variance was performed for each of the therapist rating scale items. The results indicate that for perceptions of expertness, there were significant differences as a result of treatment; $F(1, 96) = 5.54, p < .05$, with the assertive training therapist perceived as more expert. There were no significant differences on understanding. For appeal there was a significant treatment effect, $F(1, 96) = 6.07, p < .05$; the assertive training therapist was rated as more appealing. And the estimates of the effectiveness of the treatment yielded a significant treatment effect, $F(1, 96) = 5.68, p < .05$, with the effectiveness of assertive training rated as higher than the facilitative conditions. There were no significant differences as a result of sex or the Treatment \times Sex interaction.

To examine the perceptions of clients, the study contained 61 male clients of the Veterans Administration Hospital who were either outpatients at a mental hygiene clinic or patients who were being discharged from an acute inpatient treatment facility. Diagnoses were primarily personality disorders or mixed neuroses. The criteria for inclusion were a willingness to participate, being under 35 years of age, involvement in individual therapy, being considered ready for discharge, and being judged as free of observable symptoms by the therapist involved and two other staff persons making independent judgments. Five clients declined to participate, and six were rated by one or more of the judges as not being free of observable symptoms. The final sample, gathered over an 18-month period, was composed of 50 clients. The mean age was 28.6 years, with a range of 20-35; and the mean educational level was 12.93 years, with a range of 10-18 years. There were no significant differences between the groups on the basis of age or educational level. The transcripts and procedures were the same as for the therapists.

The mean ratings were slightly to moderately positive. A one-way analysis of variance revealed no significant treatment differences on any of the rating scale items.

Overall, the results for the therapist group indicate a preference for the treatment that combined the facilitative conditions with the suggestion of assertive training. The lack of significant differences on understanding would appear to be the result of the higher ratings that the therapists in the facilitative conditions treatment gave to this variable relative to the others.

The lack of significant differences in the rat-

ings of the client group can be seen as countering the expectation that this group would be likely to prefer the suggestion of a specific treatment in a manner analogous to the stereotypical physician-patient interaction. It may be that the client group, having experienced therapy, had a greater appreciation of the facilitative responses. A related possibility is that clients in therapy may have been taught to not expect specific suggestions from their therapists.

An alternative explanation may be that the client group simply lacked the background in experience and theory that the therapist group had and used in rating the transcripts. Similarly, it may be that the particular efforts made to provide valid facilitative responses and an appropriate treatment had a greater effect on the therapists than on the clients. In any event it seems reasonable that the expectations of clients about appropriate treatments were less clear than those of therapists. That the client group rated both treatments positively can be seen as an indication of an overall receptiveness to psychotherapy.

The previous study that used the same transcripts (Slaney, 1977) contained students in introductory psychology courses. The results for this group were similar to the therapist results. Although the reasons for this similarity are not clear, the differences between the student group and the client group raise questions about extending results of studies containing college students to clients, particularly clients in inpatient or outpatient treatment settings.

The lack of women clients is, of course, an important limitation of this study. Another important qualification is that the transcripts presented only one particular client and two possible approaches to treatment. How or whether the perceptions of treatment for counselors and clients vary as a function of the problem that is presented or the treatment suggested is a question that will require further research.

References

- Bergin, A. E., & Suinn, R. M. Individual psychotherapy and behavior therapy. *Annual Review of Psychology*, 1975, 26, 503-556.
- Carkhuff, R. R. *Helping and human relations* (Vol. 2). New York: Holt, Rinehart & Winston, 1969.
- Lazarus, A. A. *Behavior therapy and beyond*. New York: McGraw-Hill, 1971.
- Slaney, R. B. Perceptions of alternative roles for the facilitative conditions. *Journal of Counseling Psychology*, 1977, 24, 169-172.

Received May 25, 1977 ■

Positive Versus Negative Self-Monitoring in the Self-Control of Smoking

David A. Kantorowitz, Joyce Walters, and Kathy Pezdek
California State College, San Bernardino

This experiment compared smoking treatment programs using negative (recording number of cigarettes smoked) versus positive (recording number of urges resisted) self-monitoring. Subjects participated in one of two similar broad-spectrum treatment programs, within which they either used positive or negative self-monitoring. Over treatment, subjects in both self-monitoring groups demonstrated similarly significant reductions in smoking frequency as compared with a no-treatment control group. These findings were generally maintained at follow-up. Clinical findings call into question the accuracy of the "positive" and "negative" labels used to designate the two self-monitoring modes.

Clinical programs for smoking reduction are generally assessed by subject self-monitoring (Thoresen & Mahoney, 1974). Although several points prior to, during, and after the act of smoking would appear monitorable, clinical researchers have generally asked subjects to record their (a) successfully resisted urges to smoke (positive self-monitoring) or (b) number of cigarettes smoked (negative self-monitoring) (McFall & Hammen, 1971).

Thoresen and Mahoney (1974) have suggested that monitoring resisted urges may prove to be more facilitative to smoking reduction than monitoring the number of smoked cigarettes. Presumably, the act of self-monitoring successfully resisted urges becomes secondarily reinforcing and transforms the urges into SDs for self-control. By focusing on number of cigarettes smoked, SDs for self-punishment may be created, but since these cues appear after the behavior to be controlled, they may be less helpful in controlling the preceding smoking behavior.

Previous research has attempted to assess the effects of self-monitoring free of confounding treatment techniques. McFall (1970) reported that negative monitoring of "unmotivated" students increased cigarette consumption over a 13-day period, whereas positive monitoring decreased consumption. Working with "motivated" students,

however, McFall and Hammen (1971) found similarly significant decreases in smoking among positive and negative monitoring groups. In addition, positive and negative monitoring led to significant changes in study time (Johnson & White, 1971) and in nail-biting (McNamara, 1972), with no significant differences in frequency of smoking between the two modes.

A methodological problem with the aforementioned studies is the confounding of self-monitoring methods with subsequent, uncontrolled usage by motivated subjects of self-evaluation and covert self-praise or punishment (Thoresen & Mahoney, 1974). Additionally, there exists no research comparing different monitoring methods as the assessment vehicle for clinically relevant broad-spectrum self-control programs. This would appear to be an important area in view of the transitory nature of treatment effects produced by self-monitoring alone (Kazdin, 1974). The present study was thus designed to compare positive versus negative self-monitoring as each interacted with similar broad-spectrum behavioral self-control programs for the reduction of cigarette smoking.

Nine volunteer subjects (M age = 36.7) were assigned to self-control with positive monitoring, self-control with negative monitoring, and waiting list control groups. Treatment subjects were randomly assigned to either the positive or negative monitoring groups; subjects who could not attend either of the treatment groups due to scheduling restraints were assigned to the control group. The baseline smoking rates of the positive monitoring, negative monitoring, and control groups were 19.7, 26.8, and 15.6, respectively.

This study is based on a master's thesis completed by the second author, submitted to California State College, San Bernardino.

Requests for reprints should be sent to David A. Kantorowitz, Department of Psychology, California State College, San Bernardino, California 92407.

Both self-control treatments consisted of eight 90-minute group meetings distributed twice a week over 4 weeks. To insure that both experimental groups were taught similar techniques of self-control, each of the treatment sessions was prearranged and standardized. Both groups were instructed to (a) identify, avoid, or isolate cues for smoking, (b) make use of incompatible responses when feeling urges to smoke (such as relaxing, chewing, sucking cloves, etc.), (c) use self-talk and imagery as vehicles for self-reward and punishment, and (d) write contingency contracts for themselves and with others. Subjects in both groups were instructed to reduce their cigarette consumption at their own rate but to attempt to reach abstinence by the eighth session.

Subjects in the negative self-monitoring group were instructed to advance their counters each time they yielded to an urge and decided to smoke. Subjects in the positive self-monitoring group advanced their counters each time they resisted an urge to smoke. Both groups transcribed their daily totals onto written charts, which they brought to treatment sessions.

The control group decreased over treatment by a mean of 1.1 cigarettes a day; one subject reached abstinence. The positive and negative monitoring groups decreased by 14.7 (74.6% of baseline) and 16.8 (62.8%) cigarettes a day, respectively. The number of subjects reaching abstinence were four (positive monitoring) and two (negative monitoring). The follow-up rates of the negative and positive monitoring groups remained, respectively, 16.0 and 10.1 below the baseline of each group; three positive and three negative monitoring subjects remained abstinent.

Because the variance among groups on the pretreatment baseline measure was nonhomogeneous, $F_{\max}(3, 17) = 7.28$, $p < .05$, Tukey's *a posteriori* comparisons among means were made on each of the three treatment groups separately. There was no change in smoking frequency over time for the control group. The negative monitoring group, however, significantly reduced its smoking frequency from baseline to end of treatment ($q = 6.48$, $p < .01$) and from baseline to follow-up ($q = 6.17$, $p < .01$). The positive monitoring group significantly reduced its smoking frequency from baseline to end of treatment ($q = 5.68$, $p < .01$); its smoking frequency was suggestively, but not significantly, different from baseline to follow-up ($q = 3.89$). The nonsignificant finding was due to the highly deviant smoking rate ($z = 5.62$) of one subject. The remaining eight subjects in the positive monitoring group

maintained a significant reduction in smoking at follow-up ($q = 5.35$, $p < .05$).

t tests were performed on the relative differences between baseline and end of treatment, and baseline and follow-up, for the positive versus the negative self-monitoring groups. Contrary to prediction, no significant difference was found between the positive monitoring group and the negative monitoring group at end of treatment or at follow-up. A chi-square analysis of the number of subjects in each group who reached smoking abstinence indicated no difference between positive and negative monitoring at end of treatment or at follow-up. There was a significant difference, however, between the combined self-control groups as compared with the control group at end of treatment and at follow-up, $\chi^2(1) = 5.39$, $p < .025$, at both times.

Since both monitoring groups were paired with broad-spectrum self-control programs, it is not possible to conclude that the results support a facilitative effect of self-monitoring alone. The data do indicate, however, that contrary to the speculations of Thoresen and Mahoney (1974), there was no significant difference in outcome between positive and negative self-monitoring when used as the assessment vehicle for a broad-spectrum self-control program.

Interestingly, only one subject in the negative self-monitoring group reported that the self-monitoring treatment was "negative" or self-punishing on a follow-up questionnaire. Most subjects tended to view negative monitoring in the overall context of their treatment goals, frequently citing the self-reinforcing aspects of watching unresisted urges decline. Unexpectedly, five of eight positive monitoring subjects who completed the follow-up questionnaire indicated dislike of and frustration with this method even though their frequency of resisted urges did increase. Generally cited reasons included "preferring to see cigarette consumption decline" or "lacking feelings of accomplishment." Rather than stressing self-reinforcing effects pursuant to resisting urges, positive monitoring subjects stressed heightened frustration due to increased awareness of resisted urges to smoke. These results suggest that "positively" and "negatively" designated self-monitoring modes may be misnomers. Self-presentation of praise or punishment may solely be a function of the desirability of the self-monitored feedback that is received and independent of the method by which this feedback is procured.

Additional research would be required to investigate whether these speculations match sub-

jects' subjective appraisal patterns subsequent to positively or negatively monitoring clinical progress or reversals

References

- Johnson, S. M. & White, G. Self-observation as an agent of behavioral change. *Behavior Therapy*, 1971, 2, 488-497.
- Kazdin, A. E. Self-monitoring and behavior change. In M. J. Mahoney & C. E. Thoresen (Eds.), *Self-control: Power to the person*. Monterey, Calif.: Brooks/Cole, 1974.
- McFall, R. M. The effects of self-monitoring on

- normal smoking behavior. *Journal of Consulting and Clinical Psychology*, 1970, 35, 135-142.
- McFall, R. M., & Hammen, C. L. Motivation, structure, and self-monitoring: The role of nonspecific factors in smoking reduction. *Journal of Consulting and Clinical Psychology*, 1971, 37, 80-86.
- McNamara, J. R. The use of self-monitoring techniques to treat nail biting. *Behaviour Research and Therapy*, 1972, 10, 193-194.
- Thoresen, C. E., & Mahoney, M. J. *Behavioral self-control*. New York: Holt, Rinehart & Winston, 1974.

Received June 6, 1977 ■

Aspects of the Self-Concept Related to Level of Self-Esteem

Kenneth W. Christian
University of California at Davis

Aspects of the phenomenal self-concept of 30 male subjects varying in self-esteem, using a numerical self-report approach, were studied. Numerical ratings of *importance* and *salience* of self-enumerated positive and negative characteristics were used to generate a series of scores. Significant differences ($p < .005$) were found on an overall self-esteem score, which correlated .59 ($p < .001$) with an independent measure. No significant differences were found on ratings of positive characteristics. Striking differences were noted for negative characteristics. Results suggest that it is how individuals experience negative rather than positive characteristics that plays a determining role in self-esteem.

Difficulties involved in directly assessing the phenomenal self-concept (Wyllie, 1974) have caused recent attention on indirect assessment devices and alternative conceptualizations of self-esteem (e.g. Coopersmith, 1967; Gergen, 1971; Hagey, Smith, & Long, 1969). However, sparse data regarding regularities in the experienced self-concept would amplify our understanding of self-esteem, the present study attempts to examine aspects of the phenomenal self-concept of subjects varying in self-esteem by using a numerical self-report (NSR) approach.

The NSR approach (Christian, 1973) involves attaching subjective meanings to the two extremes of a 100-point scale and requiring subjects to respond numerically regarding this dimension of experience. In the present study, 30 male volunteers, 10 from each of three self-esteem levels (i.e., high, medium, and low) on Coopersmith's Self-Esteem Inventory (SEI), were randomly assigned to interview times with a male interviewer. After an introductory period, the subject was asked to enumerate his positive characteristics. NSRs of the *importance* of each characteristic and ratings for the *salience* (relative presence) of each were requested. Negative characteristics were then elicited, and the same rating process was repeated.

Group differences on an overall self-esteem score generated from the various NSRs were sig-

nificant at the .005 level, $F(2, 27) = 7.898$, and means were in the predicted direction. Furthermore, subjects' scores on overall self-esteem correlated .59 ($p < .001$) with scores on the SEI. These results appear to indicate that as predicted, NSRs concerning various aspects of the self-concept can be combined and transformed into an overall score that discriminates between subjects with different self-esteem levels.

Even though none of the analyses of the various subscores for positive aspects of the self were significant, significant differences were found on the negative self-regard score, $F(2, 27) = 10.244$, $p < .001$, on saliency ratings for negative characteristics, $F(2, 27) = 6.652$, $p < .01$, and on the number of negative characteristics mentioned, $F(2, 27) = 3.806$, $p < .05$.

The high-self-esteem subjects' low negative ratings seem to indicate that they do not experience their negative characteristics in the same way that others do, and that the important difference between individuals differing in self-esteem is the amount of negative, rather than positive, self-regard that they experience.

The pattern of results for the medium group is noteworthy, since they frequently scored highest on both negative and positive measures. These findings may indicate ambivalence and/or uncertainty.

References

- Christian, K. A method of assessing self-esteem through numerical self-reports (Doctoral dissertation, University of California at Davis, 1972).

This study was conducted while the author was lecturer at the University of California at Davis. Requests for reprints and for an extended report on this study should be sent to Kenneth W. Christian, who is now at the Lafayette Therapy Center, 936 Irving Avenue, Suite G, Lafayette, California 94549.

- Dissertation Abstracts International*, 1973, 34, 829H-A70H (University Microfilms No 73-19076)
- Crosger-Smith S. *The antecedents of self-esteem*. San Francisco: Freeman, 1967.
- Gergen K J. *The concept of self*. New York: Holt, Rinehart & Winston, 1971.
- Wylie R C. *The self concept* (Vol 1) *A review of methodological considerations and measuring instruments*. Lincoln: University of Nebraska Press, 1974.
- Zaller R C, Hagey J, Smith M D C, & Long B H. Self-esteem: A self-social construct. *Journal of Consulting and Clinical Psychology*, 1969, 33, 84-95.

Received June 17, 1977 ■

Changes in Preferences for Male and Female Counselors

Elaine F. Walker and Jayne E. Stake

University of Missouri—St. Louis

Past studies have indicated a preference for male therapists among both client and nonclient samples. To test for possible changes in preferences since the time of these studies, 53 male and 76 female applicants for counseling and 140 male and 150 female nonclient undergraduates completed a university counseling service application form that included a question regarding preference for sex of therapist. Although more clients than nonclients expressed preferences, results from both groups indicated a decrease in male counselor preferences.

Past studies of preferences for counselors have indicated that more subjects preferred a male therapist than a female therapist. This finding has been found for both male and female clients (Boulware & Holmes, 1970; Fuller, 1964) and male and female nonclients (Fuller, 1964). One exception to this rule has been the tendency of female nonclient subjects to indicate a preference for a female counselor for some counseling problems (Fuller, 1964).

Boulware and Holmes (1970) reported that their subjects expected male therapists to be more empathic, more knowledgeable, more experienced, and better adjusted; and, presumably, these expectations were the basis for male preferences. However, recent sex role research suggests that previously held notions about the superior competence of male professionals are disappearing (Chobot, Goldberg, Abramson, & Abramson, 1974; Levenson, Burford, Bonno, & Davis, 1975). If this is the case, then a decrease in preferences for male counselors may be expected. The purpose of this study was to explore possible changes in client and nonclient preferences for male and female counselors.

The subjects were 290 undergraduates enrolled in introductory social science courses (140 males and 150 females) and 129 applicants to a university counseling service (53 males and 76 females). Subjects were predominantly from a white, middle-class urban-suburban background. Clients were asked to state their counselor preferences as a part of their applications to a counseling service. As in Fuller's study (Fuller, 1964), subjects could indicate a male preference,

a female preference, or no preference. Nonclient subjects were tested in groups of 15-35 by either a male or female experimenter. They were asked to fill out portions of the application form completed by the clients as though they were applying for counseling.

The percentages of client and nonclient subjects preferring a male or female therapist and the percentages having no preferences are presented in Table 1.

Two questions regarding these sex preferences were considered. First, were subjects more likely to have a preference or not? Second, of those subjects who did state a preference, were they more likely to state a preference for a male or for a female counselor?

The number of male nonclients stating a preference was significantly smaller than the number stating no preference, $\chi^2(1) = 28.35, p < .001$; the number stating a preference for a male counselor was significantly smaller than the number stating a female preference, $\chi^2(1) = 7.61, p < .01$. In the case of female nonclients, the number stating a preference was significantly smaller than the number stating no preference, $\chi^2(1) = 21.66, p < .001$; among those who did state a preference, an approximately equal number preferred each sex. In summary, over 70% of the nonclients gave no preference, and a preference for a male counselor was given by only 12% of the total group.

To determine if the presence of female experimenters in this study accounted for differences between the results of this and earlier studies (which included male experimenters only), the relationship between sex of experimenter and preference (male, female, or none) was tested by a chi-square test of independence. This relationship was not significant; hence, the

Requests for reprints should be sent to Jayne E. Stake, Department of Psychology, University of Missouri, St. Louis, Missouri 63121.

Table 1
*Percentages of Nonclients and Clients
 Expressing Preferences for Counselors*

Subjects	Counselor preference			n
	Male	Female	None	
Nonclient				
Male	7.14	20.00	72.86	140
Female	16.00	14.67	69.33	150
Client				
Male	18.87	16.98	64.15	53
Female	3.95	46.05	50.00	76

decrease in preference for males cannot be explained by the presence of female experimenters.

Although male clients tended not to give a preference, the difference in number giving and not giving a preference was not significant; among males with a preference, an approximately equal number stated a preference for each sex. Among female clients, half stated a preference, and more of these subjects preferred a female counselor, $\chi^2(1) = 25.29$, $p < .001$. These findings also indicate a decrease in preference for males; in the total client sample, only 10% preferred a male counselor.

Differences in the preferences of client and nonclient groups were tested in 2×2 chi-square analyses. The client group was more likely to state a sex preference than was the nonclient group, $\chi^2(1) = 29.26$, $p < .001$; among those stating a preference, clients were more likely than nonclients to state a preference for a same-sex counselor, $\chi^2(1) = 21.89$, $p < .001$. Hence, the client group did tend to express more concern for the sex of their counselors than did the nonclient group.

These differences between the client and nonclient samples may reflect the different circumstances under which preferences were given, or

they may reflect a difference in the type of undergraduates included in the client and nonclient groups.

Although the preferences of the nonclient sample do not coincide closely with the preferences of the perspective clients, the nonclients in the present study can be compared to nonclients in earlier studies (Boulware & Holmes, 1970; Fuller, 1964). Fewer nonclients in the present study indicated a preference for male counselors. This finding suggests that most present-day undergraduates do not view male counselors as superior to female counselors.

The client group is more clearly representative of applicants for counseling. On the basis of the client sample, it appears that a substantial proportion of prospective clients do have preferences; however, in contrast to earlier findings, most of the female clients who indicated a preference wished to see a female counselor, and almost half of the male clients also preferred a female counselor. This change suggests that female counselors are now being viewed more positively by applicants for counseling.

References

- Boulware, D. W., & Holmes, D. S. Preferences for therapists and related expectancies. *Journal of Consulting and Clinical Psychology*, 1970, 35, 269-277.
- Chobot, D. S., Goldberg, P. A., Abramson, L. M., & Abramson, P. R. Prejudice against women: A replication and extension. *Psychological Reports*, 1974, 35, 478.
- Fuller, F. F. Preferences for male and female counselors. *Personnel and Guidance Journal*, 1964, 42, 463-467.
- Levenson, H., Burford, B., Bonno, B., & Davis, L. Are women still prejudiced against women? A replication and extension of Goldberg's study. *Journal of Psychology*, 1975, 89, 67-71.

Received June 21, 1977 ■

Imaging Vividness and the Outcome of in Vivo and Imagined Scene Desensitization

John M. Dyckman and Philip A. Cowan
University of California, Berkeley

This study reexamined the role of imaging vividness in desensitization success. Scores on the Betts Questionnaire on Mental Imagery were used to divide 48 snake-phobic subjects into high, medium, and low vivid groups, who were assigned to imagined scene or in vivo desensitization treatments. Imaging vividness was assessed at scheduled points during therapy. Significant decreases in behavioral and self-reported fear were observed after both treatments, though in vivo desensitization produced significantly greater fear reduction. Intherapy imaging vividness scores were significantly correlated with therapeutic success and were superior to pretherapy ratings as predictors of outcome.

Imagination of fear-relevant scenes is central to most systematic desensitization procedures. However, previous investigators have failed to find a positive relation between imaging vividness and desensitization success. To examine this apparent paradox, we refined the method of previous studies and conducted the following experiment.

Forty-eight snake phobics were pretested to establish initial fear level. A behavioral avoidance test, self-ratings of fear during this test, a questionnaire on attitudes toward snakes were used to assess fear of snakes, and the Fear Survey Schedule (Wolpe & Lang, 1964) was used to measure general fearfulness. Subjects were then separated into high, medium, and low imaging ability groups on the basis of scores on the Betts Questionnaire on imagery vividness, and were assigned to one of two standardized, individually administered desensitization procedures: imagined scene (conventional) or in vivo desensitization. Subjects in in vivo treatment were asked to enact, rather than imagine, each hierarchy scene. Imaging vividness during therapy was assessed by subject self-ratings at 19 scheduled points. Posttesting was scheduled between 1 and 3 days after completion of the treatment, and the behavioral and self-report measures were repeated.

Analyses of variance of Treatment \times Pretherapy Imaging Vividness Condition indicated no

significant differences among the groups on initial fear levels. Repeated measures analyses of variance revealed that mean posttherapy levels of fear were significantly lower than at pretesting on all measures. Snake phobics did not show significant improvement on repeated testing alone (McLemore, 1972), so the observed changes may be attributed to the treatment.

Analyses of covariance with pretherapy levels of the dependent measures as covariates showed that in vivo therapy produced significantly greater mean fear reduction than imagined scene desensitization on all measures except the Fear Survey Schedule. In vivo treatment was also a more stressful experience for the subjects, who required on the average more scene presentations to complete the hierarchy (for in vivo, $M = 81.1$, for imagined scene, $M = 79.4$) $t(46) = 2.24$, $p < .05$.

In imagined scene desensitization pretherapy, imaging vividness scores correlated $-.41$ ($p < .05$) with posttherapy avoidance of a live snake but nonsignificantly with all other measures. Imaging vividness measured during therapy, however, was highly associated with improvement, correlating with posttherapy levels of the dependent variables as follows: $-.70$ ($p < .001$) with behavioral avoidance of a live snake, $-.60$ ($p < .01$) with self-rated fear during the behavioral test, $-.49$ ($p < .05$) with attitudes toward snakes, and $-.02$ (ns) with the Fear Survey Schedule.

These results suggest several conclusions. Pretherapy imagery inventories like the Betts are currently of limited utility in predicting desensitization success, but imaging performance is

Requests for reprints and for an extended report of this study should be sent to John M. Dyckman, who is now at the Psychiatry Clinic, Kaiser Medical Center, 975 Sereno Drive, Vallejo, California 94590.

important. Betts scores are useful in screening out very low vivid imagers (in our sample, those who scored above 102.5). These patients can benefit greatly from the extra time and expense required to arrange in vivo treatment. The Betts, or any pretherapy imaging measure, should include items taken directly from the proposed hierarchy to assess the subject's ability to clearly imagine stressful material. Imaging ability is clearly related to the outcome of imagined scene desensitization. The failure of previous studies to disclose this relation may stem from a failure to assess imaging activity during

treatment, and it reminds us that psychotherapy research must combine examination of process with evaluation of outcome to produce meaningful results.

References

- McLemore, C. W. Imagery in desensitization, *Behaviour Research and Therapy*, 1972, 10, 51-57.
Wolpe, J., & Lang, P. J. A fear survey schedule for use in behaviour therapy. *Behaviour Research and Therapy*, 1964, 2, 27-30.

Received July 26, 1977 ■

Psychophysiology of Fear Imagery: Differences Between Focal Phobia and Social Performance Anxiety

Theodore C. Weerts and Peter J. Lang
University of Wisconsin

Spider phobics and speech anxious subjects imaged fear scenes with spider and public-speaking content and a series of standard scenes that were constructed to vary in degree of emotional arousal and movement. Heart rate, skin conductance, and ocular activity were recorded. Spider phobics rated all imagery contents as more vivid and reported more scene movement than speech anxious subjects. Both groups responded to their own fear scenes with higher ratings of emotion and a greater physiological response than to the other group's fear scenes. The arousal response of spider phobics to relevant fear scenes was greater than that of speech anxious subjects. The data suggest that the outcome of imagery-based therapies may be partly determined by type of fear.

The present research reexamined the finding (Lang, Melamed, & Hart, 1970) that small animal phobics and subjects with anxiety over public speaking differ in the vividness and intensity of their emotional imagery. Specifically, it was designed to determine if focal phobics give higher vividness and emotional arousal ratings than subjects with performance anxiety, are more responsive physiologically irrespective of scene content, or whether they show this greater response only to emotionally activating or focal fear materials. It was also designed to examine projected visual movement in fear imagery, as rapidity of stimulus change is associated with the genesis and evocation of small animal fears and to consider trait anxiety and general physiological reactivity as factors prompting imagery differences between fear groups.

Method

Two groups of undergraduate subjects were studied: 9 females and 2 males with a maximum spider item response on the Fear Survey Schedule (FSS), and with Spider questionnaire scores at the 75th percentile or above (all of these subjects professed no speech anxiety on the FSS and were below the Speech Anxiety Questionnaire

median); 13 females and 3 males fearful of public speaking on the FSS and from the top 25% of students in public-speaking fear (reporting no spider fear (FSS) and below the median on the Spider Fear Questionnaire (Klorman, Weerts, Hastings, Melamed, & Lang, 1974)).

All subjects visualized eight standard scenes, two from each of four content categories: scenes (a) high in emotional arousal and high in movement (e.g., panic in a burning movie theater); (b) high in movement and low in arousal; (c) low in movement and high in arousal; and (d) low in both movement and arousal (e.g., sitting at a table in the library). All subjects were also administered four fear scenes, two relevant to spider phobia (e.g., a spider crawling up your sleeve) and two related to performance anxiety (e.g., presenting a report in class).

The scenes were presented as text on a computer oscilloscope display with a 30-sec eyes-closed period for visualization. After each scene subjects reported emotional arousal during the image, perceived movement, and image vividness.

During all scenes, horizontal and vertical eye movements, skin conductance, and heart rate (HR) were continuously monitored. In addition, spontaneous skin conductance activity and habituation to a 100-dB (A) tone were assessed.

Results

The two fear groups did not differ in Taylor Manifest Anxiety Scale (TMAS) scores, in spontaneous skin conductance activity, or in skin conductance response (SCR) habituation to pure tones.

This research was supported in part by National Institute of Mental Health Grant 10993. Requests for reprints and an extended report of this study should be sent to Peter J. Lang, Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706.

Standard Scene Ratings

There was no significant group difference in report of emotional arousal to the standard scenes ($F < 1$). However, spider phobics gave significantly higher vividness ratings than did speech anxious subjects to all scenes, $F(1, 25) = 7.47$, $p < .02$. Spider phobics also tended to see more movement in the standard scenes than subjects with performance anxiety, $F(1, 25) = 3.63$, $p < .10$.

All subjects found the low-arousal scenes to be less emotionally stimulating, $F(1, 25) = 99.36$, $p < .001$, and marginally more vivid than the high-arousal scenes, $F(1, 25) = 4.14$, $p < .06$. The highest vividness ratings were generated by high-movement scenes when the arousal properties of the scene were low. High movement in combination with high arousal produced the lowest vividness ratings, $F(1, 25) = 7.00$, $p < .03$. The latter findings are consonant with Lang's (1977) view that multiple response propositions increase the information-processing load in imagery and could thus attenuate reported vividness.

High-TMAS subjects regardless of their specific fear tended to give similar vividness ratings to all scenes; low-anxious subjects showed a specific reduction in vividness reports to high-arousal scenes, $F(1, 32) = 6.07$, $p < .025$. This finding suggests that the lower mean vividness ratings found for arousing stimuli than for neutral stimuli in the total sample may be due to the low-anxiety subjects, who are less likely to spontaneously image arousal contents and are thus less practiced in this task.

Physiological Response to Standard Scenes

No simple differences between fear groups in response to the standard scenes were observed for any physiological measure. However, for the spider phobics high arousal was associated with an increase in eye movement, whereas for the speech anxious, low-arousal scenes generated more eye movement, $F(1, 23) = 5.56$, $p < .05$ (vertical movement). Also, speech anxious subjects, unlike spider phobics, increased HR during high-movement scenes, $F(1, 22) = 6.71$, $p < .03$. Although these findings imply some physiological difference between fear groups, the interactions were not predicted, and cross-validation should precede further interpretation.

Fear Scene Ratings

The four fear scenes were administered randomly among the standard scenes but were subjected to a separate statistical analysis. The spider phobics rated fear scenes as more vivid than did public-speaking anxious subjects, $F(1, 25) = 13.74$, $p < .05$. The spider phobics also rated spider scenes as significantly more vivid than the public-speaking scenes; the two fear scene types did not differ in vividness for the speech anxious, and both their ratings were similar in level to the ratings assigned by spider phobics to speech anxiety scenes, $F(1, 25) = 5.54$, $p < .05$.

Spider phobics tended to rate both fear-relevant and fear-irrelevant scenes as more arousing than did speech anxious subjects, $F(1, 25) = 12.53$, $p < .005$. However, rating differences between the two scene types were about equal for the two fear groups, and both groups assigned emotional arousal ratings to their fear-irrelevant scenes similar to that given to their standard low-arousal scenes.

Spider phobics found significantly more movement in their fear-relevant scenes than they did in the speech scenes, and they found more in these spider scenes than the speech anxious reported experiencing in either scene type, $F(1, 25) = 15.67$, $p < .001$.

Physiological Response to Fear Scenes

Eye movements did not vary significantly with either scene type or phobic group. However, both groups tended to produce faster HR, $F(1, 23) = 7.34$, $p < .03$, and a greater total SCR, $F(1, 25) = 7.40$, $p < .025$, to their fear-relevant scenes than to fear-irrelevant scenes. Furthermore, the spider phobics had higher mean responses in their relevant fear imagery (HR = 78.3 beats/min, SCR = $.015 \mu\Omega^{-1}$) than the speech anxious in their relevant scenes (HR = 74.5 beats/min, SCR = $.008 \mu\Omega^{-1}$).

Discussion

In summary, the data support the hypothesis that small animal phobics generate more vivid imagery than subjects with social performance anxiety. The fear-relevant images of spider phobics are not only more vivid, but they also prompt somewhat stronger physiological response and higher ratings of affect. The results

are also consistent with the notion that object movement may play a role in small animal fears. These differences between fear groups were unrelated to differences in anxiety or general patterns of physiological reactivity. They are consistent with the hypothesis that focal phobics' imagery, experiences of fear situations, and perhaps their potential therapeutic responses, are significantly different from those of socially anxious subjects. These results prompt us to re-evaluate research on imagery therapy that was based on only one of these populations.

References

- Klorman, R., Weerts, T. C., Hastings, J. E., Melamed, B. G., & Lang, P. J. Psychometric description of some specific fear questionnaires. *Behavior Therapy*, 1974, 5, 401-409.
- Lang, P. J. Imagery in therapy: An information processing analysis of fear. *Behavior Therapy*, 1977, 8, 862-886.
- Lang, P. J., Melamed, B. G., & Hart, J. D. A psychophysiological analysis of fear modification using an automated desensitization procedure. *Journal of Abnormal Psychology*, 1970, 76, 220-234.

Received July 29, 1977 ■

Correlations and Factor Analysis of the WISC-R and the Peabody Picture Vocabulary Test for an Adolescent Psychiatric Sample

Alan Dellhorn

Department of Psychiatry
Hennepin County Hospital, Detroit, Michigan

Valerie Klinge

Lafayette Clinic, Detroit, Michigan

Wechsler Intelligence Scale for Children Revised (WISC-R) and Peabody Picture Vocabulary Test (PPVT) scores of adolescent psychiatric patients were correlated and factor analyzed. The purposes were to explore the factor structure of these scores among a psychiatric sample, the comparability with other samples in the literature and to determine whether the PPVT adds *independent* information to an intellectual assessment. Three factors emerged that were similar to previously reported factors among other adolescent samples. The PPVT loaded heavily on the factors emerging from the WISC-R, suggesting that the PPVT adds little to information gained from the WISC-R.

The purposes of this study were to explore the factor structure of the Wechsler Intelligence Scale for Children Revised (WISC-R) with an adolescent psychiatric sample and to compare the factor structure of this sample with those previously reported. Kaufman (1971) and Kaufman & Kaufman (1975) reported that the WISC-R and PPVT scores among the adolescent psychiatric sample were similar to those reported for the general adolescent population. The WISC-R and PPVT scores were correlated and factor analyzed. The results of the factor analysis of the WISC-R and PPVT scores were compared with the results of the factor analysis of the WISC-R and PPVT scores reported by Kaufman & Kaufman (1975).

The WISC-R and PPVT scores were correlated and factor analyzed. The results of the factor analysis of the WISC-R and PPVT scores were compared with the results of the factor analysis of the WISC-R and PPVT scores reported by Kaufman & Kaufman (1975). The results of the factor analysis of the WISC-R and PPVT scores were compared with the results of the factor analysis of the WISC-R and PPVT scores reported by Kaufman & Kaufman (1975).

The results of the factor analysis of the WISC-R and PPVT scores were compared with the results of the factor analysis of the WISC-R and PPVT scores reported by Kaufman & Kaufman (1975).

The results of the factor analysis of the WISC-R and PPVT scores were compared with the results of the factor analysis of the WISC-R and PPVT scores reported by Kaufman & Kaufman (1975).

95.01 (SD 18.00). The difference between these two scores was significant using a *t* test for unpaired means (1999): $t = 3.70, p < .001$.

A correlational matrix was computed for the WISC-R IQ scores and subtests and the PPVT IQ. The intercorrelations among the WISC-R subtest scores were all significant ($p < .05$), except that Digit Span was not significantly related to Picture Completion ($r = .10, p > .05$) or to Picture Arrangement ($r = .15, p > .05$). The significant correlation coefficients of the WISC-R subtest scores ranged from .19 to .79. The correlation coefficients of the WISC-R scores with the PPVT-B IQ ranged from .21 to .79 and were all significant ($p < .05$, at least), including those for WISC-R Full Scale IQ ($r = .78, p < .001$), the WISC-R Verbal IQ ($r = .79, p < .001$), and the WISC-R Performance IQ ($r = .65, p < .001$).

From a principal components factor analysis using a three-factor varimax rotation (three factors emerged that accounted for 93.8% of the total variance, Table 1 presents the complete pattern of loadings on the three factors and identifies those variables that exceeded the loading criterion of .4). Factor 1, Verbal Comprehension, accounted for 38.7% of the variance in a stepwise determination. Factor 2, Perceptual Organization, accounted for an additional 8.7% of the variance. Factor 3, Comprehension and Attention, accounted for another 6.5% of the variance. As can be seen, the PPVT-B IQ was above the loading criterion on both Factors 1

Table 1
Loadings of the WISC-R Scores and PPVT IQ

Test	Factor		
	1	2	3
Information	.824	.290	.142
Similarities	.750	.447	.216
Arithmetic	.444	.395	.489
Vocabulary	.822	.310	.174
Comprehension	.649	.375	.449
Digit Span	.234	.095	.370
Picture Completion	.305	.683	-.007
Picture Arrangement	.375	.577	.122
Block Design	.368	.708	.234
Object Assembly	.197	.760	.128
Coding	.272	.302	.572
PPVT IQ	.748	.443	.048

Note. WISC-R = Wechsler Intelligence Scale for Children-Revised; PPVT = Peabody Picture Vocabulary Test. The loading criterion was .40.

The factors that emerged are very similar to those reported from other samples (Kaufman, 1975; Van Hagen & Kaufman, 1975). The first and second factors, Verbal Comprehension and Perceptual Organization, appear consistently in such studies and offer construct validity to the separation of Verbal and Performance scores on the WISC-R. The third factor, including Arithmetic and Coding, is very similar to the Freedom from Distractibility factor described elsewhere (Kaufman, 1975), which also includes Digit Span. These findings suggest that the WISC-R scores of an adolescent psychiatric

sample can be compared meaningfully to those of retarded or normal children, since there seem to be no qualitative differences in the structure of the intellectual abilities tested among the three groups.

The substantial loadings of the PPVT-B IQ on the two main factors in this study suggest that the PPVT adds little to the knowledge gained from the WISC-R. Although the correlation between the PPVT-B IQ and the WISC-R Full Scale IQ is rather high, studies using other than correlational designs indicated that the PPVT IQ should not be used as an estimate of the WISC-R IQ (see, e.g., Condit et al., 1976).

References

- Appelbaum, A., & Tuma, J. Social class and test performance: Comparative validity of the Peabody with the WISC and WISC-R for two socioeconomic groups. *Psychological Reports*, 1977, 40, 139-145.
- Condit, J., Lewandowski, D., & Saccuzzo, D. Efficiency of Peabody Picture Vocabulary in estimating WISC scores for delinquents. *Psychological Reports*, 1976, 38, 359-362.
- Kaufman, A. S. Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology*, 1975, 43, 135-147.
- Van Hagen, J., & Kaufman, A. S. Factor analysis of the WISC-R for a group of mentally retarded children and adolescents. *Journal of Consulting and Clinical Psychology*, 1975, 43, 661-667.

Received August 9, 1977 ■

Suicide and Transparency Responses on the Rorschach: A Replication

Jill Rierdan
Wellesley College

Elizabeth Lang and Sara Eddy
Human Resource Institute, Boston, Massachusetts

In 1974 Blatt and Ritzler reported that the presence of transparency and cross-sectional responses on the Rorschach differentiated individuals who completed suicide from non-suicidal individuals. This study attempted to cross validate those findings by comparing the Rorschach protocols of 14 individuals who committed suicide with those of a matched sample of individuals who had not committed suicide. In keeping with Blatt and Ritzler's results, the group of suicidal individuals gave a greater number of transparency and cross-sectional responses than did the group of non-suicidal individuals thus indicating the utility of the Rorschach in predicting suicide.

In a review of studies attempting to assess the predictive validity of the Rorschach for identifying suicidal individuals Neuringer in 1965, concluded his attempt to predict suicide using different methods of reading the Rorschach have produced equivocal results. He concluded such results are methodological confusions. In a more recent, however, Neuringer not only criticized earlier work claiming an methodological validity for the Rorschach, but also suggested that the Rorschach might emerge in the future as a predictor of suicide.

Although the validity of the Rorschach as a suicide predictor has not been demonstrated, the Rorschach is a construct that characterizes individuals in the same way. However, a firm conclusion regarding its validity is premature. The Rorschach is a projective test that has been used in a number of ways and in different settings. It is a measure of personality and of psychological relationships. It is a measure of the individual's perception of the world and of the world's perception of the individual. It is a measure of the individual's perception of the world and of the world's perception of the individual. It is a measure of the individual's perception of the world and of the world's perception of the individual.

The present study of Blatt and Ritzler (1974) was a replication of their study. They compared the Rorschach protocols of 14 individuals who committed suicide with those of a matched sample of individuals who had not committed suicide. In keeping with Blatt and Ritzler's results, the group of suicidal individuals gave a greater number of transparency and cross-sectional responses than did the group of non-suicidal individuals thus indicating the utility of the Rorschach in predicting suicide.

Correspondence should be sent to Jill Rierdan, Wellesley College, Wellesley, MA 01981.

that the former group gave a significantly greater number of transparency and translucency responses on the Rorschach than did the latter. In concluding the discussion of their investigation, Blatt and Ritzler acknowledged the need for cross-validation. The present investigation provides this needed replication.

The Rorschach protocols of 14 patients who had committed suicide after their hospitalization in a short-term intensive care facility were examined. The sample of 14 suicides contained 4 females and 10 males ranging in age from 19 to 44 ($M = 27.5$), in Wechsler Adult Intelligence Scale (WAIS IQ) from 82 to 127 ($M = 107.2$), and in number of Rorschach responses from 10 to 67 ($M = 26.0$).

From the records of patients hospitalized at the same facility, 14 controls were selected using a matched-pairs technique such that the control sample contained 4 women and 10 men, whose ages ranged from 17 to 46 ($M = 29.0$), whose WAIS IQ ranged from 86 to 123 ($M = 107.2$), and whose number of Rorschach responses ranged from 11 to 62 ($M = 25.9$). t tests for matched pairs confirmed that the suicide and control groups did not differ in age, IQ, or number of Rorschach responses. In keeping with Blatt and Ritzler (1974) no attempt was made to control for previous suicide attempts in the control group. The only criterion was that at the time of the study none of the control patients was reported to have committed suicide.

Rorschach protocols were scored for presence of transparency (e.g., "a light bulb"), translucency (e.g., "a car"), and cross-sectional (e.g.,

"X ray") responses as described by Blatt and Ritzler (1974). Additionally, responses that were not clearly transparencies but that seemed to conform to Blatt and Ritzler's interpretation of transparencies as "an attempt to establish three-dimensional representations, but without the capacity to represent volume" (p. 282)—for example, "splattered paint"—were scored as criterial. All Rorschach protocols were scored independently by two judges.

When the number of transparency, translucency, and cross-sectional responses were totaled for each patient, the suicide group showed a significantly greater number of criterial responses ($M = 3.00$) than did the control group ($M = 1.78$); one-tailed t test for matched pairs, $t(13) = 2.33$, $p < .05$. The two patient groups were not found to differ when transparencies (including translucencies) and cross-sectional responses were analyzed separately.

These results replicate, in large part, the findings of Blatt and Ritzler (1974). That Blatt and Ritzler found differences between groups when analyzing transparency (and translucency) and cross-sectional responses independently while we found significant group differences only when combining these responses does not seem critical. Both classes of responses were seen by Blatt and Ritzler as exemplifying the same formal property—an inadequate representation of volume. Our findings, in cross-validating those of Blatt and Ritzler, provide evidence that a single sign on the Rorschach may be a useful predictor of suicide.

Roth and Blatt (1974), in considering why transparencies are given primarily by suicidal individuals, reasoned that this response reflected a loss or collapse of three-dimensional representation, paralleled by a loss of self-other differentiation, leading to the self and other becoming viewed as objects of aggression. This interpretation is consistent with the findings that suicidal individuals give transparency responses on the Rorschach; however, it does not explain specifically why patients who complete suicide give transparency responses more frequently than other patients who may threaten or attempt suicide.

We suggest an alternative model for understanding the association of suicide and transparencies. This model is admittedly speculative, but it is consistent with some clinical literature and is subject to empirical validation.

Since the characteristic that most obviously distinguishes completed suicides is that they are

responsible for their own death, our approach has been to consider the meaning of death for these individuals. In this regard, Morse (1973) has described suicide-promoting fantasies. In these fantasies, according to Morse, death is conceptualized as a means for satisfying important wishes, rather than an end in itself. A presupposition of suicide-promoting fantasies is that the individual will be alive after he/she kills himself/herself. Morse suggests that even though all humans are incapable of truly conceiving of their own death, suicides are distinctive in the degree to which they accept their after-death fantasies as real.

We suggest that transparency responses on the Rorschach (and, more generally, inadequate representations of volume) can be understood in the same terms as suicides' conceptions of death. Just as death seems to be viewed as a transitional phase rather than as a definite end, so too transparencies exemplify this same penetrable quality—An object is not represented as solid and bounded but as something insubstantial or pregnable.

At present, there is a dearth of systematic research assessing the conceptions of life and death held by suicidal and nonsuicidal individuals. Empirical assessment of the hypothesis of a relationship between transparencies on the Rorschach and attitudes toward death is one clear direction for research that would facilitate an understanding of the present findings and perhaps suggest further avenues of investigation.

References

- Blatt, S. J., & Ritzler, B. A. Suicide and the representation of transparency and cross-sections on the Rorschach. *Journal of Consulting and Clinical Psychology*, 1974, 42, 280-287.
- Lester, D. Attempts to predict suicidal risk using psychological tests. *Psychological Bulletin*, 1970, 74, 1-17.
- Morse, S. J. The after-pleasure of suicide. *British Journal of Medical Psychology*, 1973, 46, 207-238.
- Neuringer, C. The Rorschach as a research device for the identification, prediction, and understanding of suicidal ideation and behavior. *Journal of Projective Techniques and Personality Assessment*, 1965, 29, 71-82.
- Neuringer, C. Suicide and the Rorschach: A rueful postscript. *Journal of Personality Assessment*, 1974, 38, 535-539.
- Roth, D., & Blatt, S. J. Spatial representations of transparency and the suicide potential. *International Journal of Psychoanalysis*, 1974, 55, 287-293.

Received September 1, 1977 ■

Premorbid Social Competence Construct Generalizability Across Ethnic Groups: Path Analyses With Two Premorbid Social Competence Components

Raymond M. Costello

Department of Psychiatry, University of Texas Health Science Center at San Antonio
and Audie Murphy Veterans Administration Hospital, San Antonio, Texas

The stability across ethnic groups of interrelations among five variables was studied using path analysis. Certain findings involving four variables were consistent with theoretical expectations for the Anglo-American group but were reversed in the Mexican-American group. Methodological and theoretical implications are presented.

The stability across ethnic groups that is, Mexican Americans and Anglo Americans of interrelations among five variables: education, age at first psychiatric hospitalization, current age at admission, symptom severity, and behavioral adjustment while hospitalized, was studied using path analysis. Subjects were 14 male Mexican Americans and 14 male Anglo Americans admitted consecutively to a Veterans Administration hospital, studied treatment, and longitudinal outcome and standard deviations for Mexican Americans and Anglo Americans respectively were education = $MD = 11$ and $SD = 1.7$ and $MD = 12$ and $SD = 1.9$; current age = $MD = 31$ and $SD = 7.1$; symptom severity = $MD = 1.0$ and $SD = 1.4$; and behavioral adjustment = $MD = 0.6$ and $SD = 0.4$. The path model for symptom severity was compared across a series of four models of common variance of symptom severity. Behavioral adjustment was measured by rating scales on first psychiatric hospitalization, and was also analyzed as a function of the MA and Anglo groups. High scores reflect a negative score with good social skills.

Interrelations of interrelations within each variable were examined through path analysis. The path model for symptom severity was compared across a series of four models of common variance of symptom severity. Behavioral adjustment was measured by rating scales on first psychiatric hospitalization, and was also analyzed as a function of the MA and Anglo groups. High scores reflect a negative score with good social skills.

Copyright 1978 by the American Psychological Association. 0893-3200/78/4405-1164\$02.75

placed in a standard causal order, and a standard score transformation for each variable for each group preceded analysis. A correlation of .30 was required for significance. Ten bivariate correlations were decomposed. Nine of 42 Mexican-American coefficients and 10 Anglo-American coefficients were $\geq .30$. Significant relationships nonreplicable across ethnic groups, were observed between education and symptom severity; education and behavioral adjustment, and first hospitalization age and symptom severity. In each case the direction of the zero-order correlation, of the direct causal effect, and of the total causal effect was reversed across ethnic groups. Differences in magnitude of correlation were noted also. Significant relationships replicable across ethnic groups were observed between education and current age and first admission age and current age. Nonsignificant relationships were observed for the remaining five variable pairs.

As education and first admission age are two components of the premorbid social competence construct (Fogel & Phillips, 1961), this study has implications for psychosocial developmental theory. A successful developmental history (i.e., higher education and older age at first psychiatric breakdown) should be associated with a good prognosis (i.e., less severe symptomatology and better behavioral adjustment while hospitalized). Theoretical expectations were confirmed for Anglo Americans as three statistically significant bivariate relations within the Anglo group were of the proper sign, and none were inconsistent with theory (see Table 1). Yet, not only were these predicted relationships not found for the Mexican-American group, they were actually reversed in sign. Better educated Anglos tended

Table 1
Decomposition of Statistically Significant Bivariate Relationships for Each Ethnic Group

Bivariate relationship	(A) Total covariation		Causal						Noncausal = A - D	
			(B) Direct		(C) Indirect		(D) = B + C			
	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA
X ₃ X ₆	.24	-.44	.36	-.42	-.12	-.02	.24	-.44	None	None
X ₃ X ₄	-.01	.30	-.07	.30	.06	.00	-.01	.30	None	None
X ₃ X ₅	.33	-.24	.34	-.31	.08	.08	.42	-.23	-.09	-.01
X ₃ X ₂	-.34	-.17	-.21	-.17	-.13	.00	-.34	-.17	None	None
X ₃ X ₁	.67	.64	.63	.64	None	None	.63	.64	.04	.00

Note: MA = Mexican American; AA = Anglo American; X₆ = education; X₄ = age at first psychiatric admission; X₅ = current age; X₂ = alcoholism symptom severity; X₁ = Behavioral adjustment.

to report fewer alcohol-related symptoms and were observed to make better verbal and social contact with other patients than did poorer educated Anglos. Anglos who were older at first psychiatric admission tended to report fewer alcohol-related symptoms on current admission than did Anglos who were younger at first admission. The obverse was true for each phenomenon within the Mexican-American group.

One can only speculate about the meaning of this nongeneralizability. In the Anglo-American group, education is obviously an additional adaptational tool. In the Mexican-American group, primarily a low-socioeconomic laboring class group, however, education may have been as much an adaptational liability as an asset, as these more educated people may have been less likely to be satisfied in laboring occupations. Or, these Mexican Americans who were in a position to acquire a more advanced education may have experienced additional sociopsychological stresses not ordinarily encountered in their ethnic groups as a result of socioeconomic/occupational upward mobility.

With regard to experimental methodology in racial ethnic comparisons, techniques such as subject matching or analysis of covariance

would be inappropriate controls of extraneous subject characteristics variance if the matching variable/covariate(s) show ethnic unreliability, as was found in the present study.

The obvious conclusion is that when predictor or criterion variables (and the theory by which they are subsumed) derive from work on ethnically homogeneous or ethnically unspecified groups are to be used in work with groups of specified ethnic composition different from the derivation group, the variables must first be studied with regard to their freedom from ethnic bias (i.e., generalizability). Failure to do so may lead to distorted findings with possible social ramifications. Second, the pathanalytic formulation has good potential as an instructional device.

References

- Ellsworth, R. B. *The MACC Behavioral Adjustment Scale: Revised 1971 manual*. Los Angeles: Western Psychological Services, 1971.
- Zigler, E. & Phillips, L. Social competence and outcome in psychiatric disorder. *Journal of Abnormal Psychology*, 1961, 63, 264-271.

Received September 12, 1977 ■

Effects of the Sex of Both Interviewer and Subject on Reported Manifest Dream Content

Russ B. Kremsdorf, Lucy J. Palladino, Douglas D. Polenz,
and Barbara J. Antista
Arizona State University

The present study examined the effects of the sex of both interviewer and subject on the reported content of dreams. Three male and three female interviewers each interviewed five male and five female subjects to elicit dream reports. In contrast to previous studies, no sex differences in the sexual content of dreams were found, although the dreams of males were more vivid, active, and aggressive. Opposite-sex pairing mobilized reports of conflict within dreams, whereas same-sex pairing increased the sexual content. These results support the hypothesis that environmental factors are influential in determining the dream content that is reported.

What is often thought to be common to waking fantasy processes and dreams is that they both involve the psychological transformation of stored information with a relative lack of attention to external input. However, Whitman, Kramer, and Haldridge (1964) maintain that the conditions present at the time of dream reporting are more potent factors in determining dream content than are dream recall processes. Although experimental sleep research has begun to study the more obvious effects of setting and presleep stress on the manifest content of dreams, relatively little attention has been given to the effects of the sex of the interviewer on the dream content that is reported.

Those studies that have investigated the sex of the experimenter in the areas other than dream research have found that this variable can influence such diverse phenomena as conformity and compliance, need affiliation, empathy, classroom learning, verbal conditioning, and experimenter bias. The sex of the interviewer has also been demonstrated to affect frankness, honesty, and enjoyment of an interview when discussing sensitive areas of communication. Such findings indicate the need to consider the sex of both participants both separately and in interaction, in areas of inquiry in which this variable has formerly been neglected.

Studies of dream content have commonly focused on the differences between the dreams of

male and female subjects. In the most well-known investigation, Hall and Van de Castle (1966) found dreams of males to be generally more active, to contain more male characters than female characters, to exhibit more aggression, and to involve more sexual content than those of females when dream diaries were submitted to a male class instructor. Similar results have been reported in the literature, though in almost every study of dream content male interviewers were used. There has been no controlled experimental attempt to assess the effects of the sex of both participants on the dream content reported.

The purpose of this experiment was to systematically examine the subsequent effects on reported dream content when male and female subjects were interviewed by either male or female interviewers. This design allowed for observing whether the differences generally reported between the dreams of male and female subjects might be related to the failure of these studies to use interviewers of both sexes. Special emphasis was given to those dream content ratings considered to reflect the most basic dimensions of the dream (Hauri, Sawyer, & Rechtschaffen, 1967) and most likely to reflect differences influenced by the sex of the participants.

The subjects were 30 male and 30 female undergraduate students enrolled in introductory psychology courses. Male subjects ranged in age from 17 to 30 years, with a mean age of 21.2 years. Female subjects ranged in age from 17 to 32 years, with a mean age of 19.9 years. Interviewers were three male and three female clinical psychology graduate students of approximately the same age and therapeutic experience.

Requests for reprints and for an extended report of this study should be sent to Russ B. Kremsdorf, Department of Psychology, Arizona State University, Tempe, Arizona 85281.

Each interviewer was randomly assigned five male and five female subjects and requested each subject to report the content of the last dream he/she could recall. After allowing the subject to complete the spontaneous dream report, the interviewer clarified any ambiguities. All interviews were audio recorded. Most of the dream rating dimensions used in this investigation were derived from the factor analytic study of dream ratings by Hauri et al. (1967). The six independent dream content dimensions were vivid fantasy, hedonic tone, active control, verbal aggression, physical aggression, and sexuality. Affect in the dream reports was also assessed by the content analysis scales developed by Gottschalk, Winget, and Gleser (1969). These dimensions were anxiety, hostility directed outward, hostility directed inward, and ambivalent hostility.

Because of the relatively good reliabilities obtained on this free-response material, all data analyses used mean ratings of the four raters. Overall, the dreams reported by male and female subjects did not differ in either the intensity or degree of sexual content. However, same-sex pairings of subjects and interviewers resulted in more sexual content, an interaction approaching significance, $F(1, 56) = 3.82, p < .10$. The dreams reported by male subjects were more vivid, $F(1, 56) = 6.76, p < .025$; active, $F(1, 56) = 4.58, p < .05$; and physically aggressive, $F(1, 56) = 4.36, p < .05$; than those of females. In addition, the dreams of males exhibited more overall subjective impact than the dreams of females, $F(1, 56) = 5.40, p < .025$. Opposite-sex pairing of subjects and interviewers resulted in dream reports that reflected greater anxiety, $F(1, 56) = 6.40, p < .025$; hostility directed inward, $F(1, 56) = 6.84, p < .025$; and ambivalent hostility, $F(1, 56) = 2.78, p < .10$.

In contrast to previous assertions, the results of this study argue against the assumption that biological differences between the sexes determine the degree of sexual content within dreams. En-

vironmental factors such as the sex of the individual to whom the dream is reported seem influential in the resulting dream report. It is difficult to determine whether the differences between these results and those of studies conducted over 20 years earlier are due to cultural changes in sex roles or to the fact that earlier studies used only male interviewers. The other differences found between the dream reports of males and females are consistent with similar comparisons made in a number of studies over the last 30 years. This does raise the question of whether some basic dream processes are somehow different between the sexes.

The results pertaining to opposite-sex pairing and dream reports support the assertion that to stimulate conflict for therapeutic purposes, one should expose the individual to a member of the opposite sex. However, these findings also suggest when sexual problems are the area of concern, pairing a client with a therapist of the same sex may facilitate discussion. Finally, the researcher of dream content and the clinician need to consider that the dream report that they collect is not immune to the effects of the setting in which it is reported.

References

- Gottschalk, L. A., Winget, C. N., & Gleser, G. C. *Manual of instructions for using the Gottschalk-Gleser Content Analysis scales: Anxiety, hostility and social alienation-personal disorganization*. Berkeley: University of California Press, 1969.
- Hall, C. S., & Van de Castle, R. L. *The content analysis of dreams*. New York: Appleton-Century-Crofts, 1966.
- Hauri, P., Sawyer, J., & Rechtschaffen, A. Dimensions of dreaming: A factored scale for rating dream reports. *Journal of Abnormal Psychology*, 1967, 72, 16-22.
- Whitman, R. M., Kramer, M., & Baldrige, B. Which dream does the patient tell? *Archives of General Psychiatry*, 1963, 8, 277-282.

Received October 21, 1977 ■

Androgyny and Self-Esteem in the Upper-Middle Class: A Replication of Spence

Kevin O'Connor, David W. Mann, and Judith M. Bardwick
University of Michigan

A middle-aged upper-middle class sample was used in a replication of Spence, Helmreich, and Stapp's study of androgyny and self-esteem in an undergraduate sample. The earlier findings were largely replicated. Self-esteem scores for the men were substantially higher than those found by Spence et al., but the earlier relationships of androgyny, masculinity, and femininity with self-esteem received support. Implications of the demographic characteristics of the sample for the generality of current sex role and sex identity research are discussed.

Recent research on androgyny has focused on the relative adjustment of men and women classified as more or less masculine, feminine, or androgynous (e.g., Bem, 1975; Kaplan & Bean, 1976). These reports indicate that the healthiest individuals are not those who score at the extreme of their sex-appropriate distributions on masculinity-femininity scales. Instead, those classified as androgynous score highest on measures of adjustment and mental health (Bem, 1975; Carlson, 1971; Heilbrun, 1976; Spence, Helmreich, & Stapp, 1975).

The generality of this finding has been limited, however, by the demographics of the subjects—typically college undergraduates—used in these studies. In the present study, upper-income, middle-aged, Caucasian, suburban homeowners were studied, and their data were compared with those reported by Spence et al. (1975) in their study of college students. Thus, our research attempted to extend the generality of the theoretically important findings reported by Spence et al. in their work on androgyny.

Traditional sex role research considers masculinity and femininity as diametrically opposed clusters of attributes. Healthy individuals were considered to be those characterized by high gender-appropriate masculinity or femininity scores. As noted, contemporary work has taken the opposite view with high self-esteem predicted for the jointly masculine and feminine

individual regardless of sex (e.g., Bem, 1977; Spence et al., 1975).

Our purpose was to see whether our more traditional, established, conservative sample would produce the pattern of sex role attribute ratings and the relationship of androgyny with self-esteem reported by Spence et al.

Specifically, Spence et al. reported findings from several sets of analyses. We were concerned with the following:

1. Spence et al. compared the mean scores of men and women on each of their measures: sex role attribute ratings—endorsement of "valued" sex-typed attributes, measures of the willingness to describe "typical" men and women in sex stereotyped terms, a self-esteem scale, and a scale of agreement with feminist attitudes toward women. They reported significant sex differences in all but the self-esteem scores. We tested for sex differences using the same measures in our sample.

2. Spence et al. correlated their measures. They reported significant positive correlations between men's masculinity scores and use of masculine stereotypes and between women's femininity scores and use of feminine stereotypes. They reported that sex-appropriate sex role attribute self-ratings tended to relate negatively to feminism for men and women, but they found strong negative correlations between sex stereotyping and feminism. Finally, they found positive correlations between sex role attribute scores and self-esteem but no relation between self-esteem and feminism. We examined these correlations in our data.

3. Spence et al. placed the men and the women into groups based on their sex role attribute scores. For example, those above the

We gratefully acknowledge Mary Ellen Colten's substantive contribution to the work reported here. Requests for reprints should be sent to Judith M. Bardwick, Department of Psychology, University of Michigan, 529 Thompson, Ann Arbor, Michigan 48109.

median on masculinity were classified as high masculine, those below, low masculine. They crossed the dichotomized masculinity and femininity groups to produce a fourfold table for each sex. The cells were high masculinity-high femininity (defining androgyny), high masculinity-low femininity (traditional male), high femininity-low masculinity (traditional female), and low femininity-low masculinity (which Spence et al. call undifferentiated). Comparing the mean self-esteem scores of the four groups for each sex, they found that for each sex self-esteem means were highest in the androgyny cell and descended in the order given above.

We applied these procedures in our data to see whether this central aspect of Spence et al.'s findings would be replicated in our strikingly different sample.

Method

Procedure. Subjects were Caucasian men ($n = 43$) and women ($n = 48$) between 40 and 50 years old. All were members of an informal subdivision association who responded to a letter sent to members requesting volunteers for the study. The mean number of children in the sample was 3.26 per family; annual income for the sample ranged from \$50,000 to well over \$100,000. Many of the women were employed outside the home; and the men were employed in professional and executive positions.

The self-administered questionnaires were left at the subjects' homes and were picked up after two days during the fall of 1975. Self-esteem was measured by the Texas Social Behavior Inventory (TSBI); sex role attribute ratings and sex role stereotyping, by the Personal Attributes Questionnaire (PAQ); and degree of feminism, by the Attitude Toward Women Scale (AWS). Order of presentation was counterbalanced. All instruments are described in Spence et al. (1975) and were used as reported there.

Results

On the whole, sex differences in the data reflected those reported by Spence. The men in our sample described themselves as significantly more masculine than did the college men, $t(289) = 2.08$, $p < .05$. Unlike Spence's men, ours were no more prone to use sex role stereotypes than the women. And, in our sample as in Spence's, the women were only slightly (insignificantly) more profeminist than the men. Regarding self-esteem, our occupationally successful men scored significantly higher than the women, $t(89) = 2.35$, $p < .05$, who scored at the same level as all of Spence's subjects regardless of sex.

Similarly, correlations generally reflected relationships consistent with those of Spence, except that femininity was not related to self-esteem among the men. Masculinity and self-esteem were significantly correlated among men and women, $r(41) = .38$ and $r(45) = .77$, $ps < .05$; femininity and self-esteem correlated only among women, $r(45) = .46$, $p < .05$.

The relationship between self-esteem and masculinity reflects the vocational achievements of these men, having clearly met and surpassed the standards of competence embodied in the Spence et al. Masculinity scale. The interpersonal skills indexed in the Femininity scale seem not to contribute to self-esteem for these men, unlike the college sample.

Thus, two questions remained: (a) Could any of our samples be classified as androgynous? and (b) if so, Would the pattern of self-esteem means reported by Spence et al. hold?

The answer to both questions was a straightforward yes. Our results were in close agreement with Spence. The androgynous men and women were highest in mean self-esteem, followed, within each sex, by masculine, feminine, and undifferentiated groups.

Overall, our data lend support to and extend the generality of the work of Spence et al. Principally, androgynous self-descriptions did occur reliably in our sample and androgynous self-descriptions predicted the highest levels of self-esteem for men and women. The high esteem and masculinity scores of the men may well reflect the broader opportunities to test their competence that are available neither to their spouses nor to college students.

References

- Bem, S. L. Sex role adaptability: One consequence of psychological androgyny. *Journal of Personality and Social Psychology*, 1975, 31, 634-643.
- Bem, S. L. On the utility of procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1977, 44, 196-205.
- Carlson, R. Sex differences in ego functioning. *Journal of Consulting and Clinical Psychology*, 1971, 37, 367-377.
- Heilbrun, A. B. Measurement of masculine and feminine sex role identities as independent dimensions. *Journal of Consulting and Clinical Psychology*, 1976, 44, 183-190.
- Kaplan, A. G., & Bean, J. P. (Eds.). *Beyond sex role stereotypes*. Boston: Little, Brown, 1976.
- Spence, J. T., Helmreich, R., & Stapp, J. Ratings of self and peers on sex role attributes and their relation to self esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 1975, 32, 29-39.

Received December 13, 1977 ■

Learned Helplessness as a Model of Depression

Special issue of the
Journal of Abnormal Psychology,
February 1978.

The issue serves as a forum to investigate the validity and adequacy of the learned helplessness model of depression in humans and includes both critical and supportive theoretical and empirical articles with comments.

1. Rowell Huesmann is guest editor. Richard A. Depue, Martin E. P. Seligman, Charles G. Costello, Robert C. Smolen, and Alexander M. Buchwald are among the contributors.

Copies of the special issue are available at \$6 each, prepaid. Discounts on bulk orders are also available. To order single copies or for more information on bulk orders write:

Subscription Department
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036

CHARLES C THOMAS • PUBLISHER

MICROCOUNSELING: Innovations in Interviewing, Counseling, Psychotherapy and Psychoeducation (2nd Ed.) by Allen E. Ivey, Univ. of Massachusetts, Amherst, and Jerry K. Worell, Univ. of Nebraska Medical Center, Omaha. Foreword by Bernard G. Guernsey, Jr. (Contributors) The new edition of this book includes updated research on the microcounseling model, chapters on the cross-cultural and cross-racial implications of interviewing and counseling, and material illustrative of how microcounseling skills are used by various types of therapists. A section on operational definitions has also been added. '78, 624 pp., 13 il., \$14.00

INTRODUCING NEUROPSYCHOLOGY: The Study of Brain and Mind by Stuart J. Freedman, Univ. College, Cardiff, Wales. The study of consciousness and other enigmas of human life are discussed in this book to explain how the mind and behavior are related to the underlying machinery of the brain. An introduction of basic methods and systems is followed by chapters on the effects of brain injury and disease, the split-brain condition, the human, the cerebral cortex, the infant brain, neurological processes in psychiatric illness, development of technology for control of the brain, and other topics. Students, practitioners and researchers in psychology are among those who will find this book to be enlightening and provocative. '78, 236 pp., 12 tables, \$13.75

BEHAVIOR THERAPY IN CLINICAL PSYCHIATRY: Decision Making, Procedure and Outcome by Ernest G. Poser, McGill Univ., Montreal, Quebec, Canada. Foreword by H. J. Eysenck. This text considers the application of behavior conditioning principles to mental health problems. It explores specific fears, social withdrawal and socially disapproved behavior. Illustrative case studies are included. '78, 244 pp., 4 il., 1 table, \$13.50

THE MALTREATED CHILD: The Maltreatment Syndrome in Children—A Medical, Psychological and Social Guide (3rd Ed.) by Vincent J. Carrozza, New York Foundling Hospital, New York City, and Douglas J. Besharov, United States National Center of Child Abuse and Neglect, Washington, D. C. Foreword by Loretto Bernard. This edition covers the problems of child maltreatment; their nature, causes and extent; and the means to deal with them. '77, 176 pp., 11 il., \$12.50

CAREER DEVELOPMENT AND COUNSELING OF WOMEN edited by L. Sunny Hansen, Univ. of Minnesota, Minneapolis, and Rita S. Rapoza, Wilson Learning Corp., Eden Prairie, Minnesota. (46 Contributors) Articles are included on psychological and environmental variables which affect female development and which have an impact on their decisions, not only about work, but also about their life roles with respect to family, leisure and community. Occupational socialization, planning and negotiating for multiple roles, assessment of career interests, cross-cultural perspectives of female career development, and other topics are examined. '78, 672 pp., 57 tables, \$24.75

NEXUS PSYCHOTHERAPY: Between Humanism and Behaviorism by Kenneth Urial Gutsch, Univ. of Southern Mississippi, Hattiesburg, and Jacob Virgil Ritenour, II, Jackson, Michigan. Foreword by Howard S. Rosenblatt and Harold H. Dawley, Jr. The common ground between humanism and behaviorism is explored as it relates to emotions, behavior, values and motivation. It is a unique approach which will provoke new understandings in psychotherapy as it presents methods by which such understandings can be implemented. '78, 196 pp., 15 il., 16 tables, \$14.75

CEREBRAL LOGIC: Solving the Problem of Mind and Brain by Charles W. Needham, Neurological Associates of Tucson, Tucson, Arizona. In its inquiry into the neurophysiology of consciousness, this book considers equally the disciplines of neurology, philosophy, and psychology. A discussion of the general nature of the mind-brain problem is followed by a look at the origin of mentalistic language and its applicability to cerebral logic. '78, 232 pp., 1 il., \$12.75

THE POWER OF POSITIVE REINFORCEMENT: A Handbook of Behavior Modification by Judith Elbert Favell, Western Carolina Center, Morganton, North Carolina. This handbook systematically surveys the entire process of behavior modification. Initial chapters explain preoperational procedures such as the specification and definition of target behavior. The remainder of the book details the actual application of behavior modification, including positive and negative reinforcement, token systems, and complex behavior training. '77, 288 pp., 1 il., 1 table, \$9.75

327 East Lawrence Avenue • Springfield • Illinois • 62717
Orders with remittance sent, on approval, postpaid

AHSEN'S EPT

Description: Ahsen's Eidetic Parents Test (EPT) elicits eidetic images of parents, in the mind of the subject, in 30 situations. Each response is a triadic experiential unit — an image(I), a somatic response(S), and a meaning(M). A precise test for localizing personality structures, somatic and emotional patterns and psychological themes in the form of images which do not change unless transformed through therapeutic

Indications:

Uses: Easily administered within one hour (brief) or longer (comprehensive), if desired. A structured interview which elicits valid developmental and personality data, also an initial diagnostic instrument or therapy process, for use with adults and children. Full instructions for psychological report writing.

"A methodological advance."

— *American Journal of Psychiatry*

"An important addition to the fear inventories and reinforcement schedules currently in use...a rapid and effective method."

— *Behavior Therapy*

"An exciting and ingenious way of getting at conflict areas and providing a rich source of material for therapeutic sessions...A new and intriguing device for personality assessment."

— *Contemporary Psychology*

"Therapeutic as well as diagnostic and fact finding." — *The International Journal of Clinical and Experimental Hypnosis*

S M

EIDETIC PARENTS TEST

- ▲ CLINICAL DIAGNOSIS
- ▲ PERSONALITY ASSESSMENT
- ▲ DEVELOPMENTAL FUNCTIONING
- ▲ SPECIFIC THERAPY IMAGES

for

➔ **BRANDON HOUSE, Inc.** For orders, write P.O. Box 240, Dept. F, Bronx, N.Y. 10471

Yes, send me Ahser Ahsen's **EIDETIC PARENTS TEST AND ANALYSIS**. I am enclosing \$17.50 plus 30¢ postage and handling (N.Y.S. residents, please add 8% tax.)

Name _____

Address _____

City _____

State _____

Zip _____

Send bill to customers only.



NEW

from the APA Insurance Trust A TOTAL MEDICAL INSURANCE PACKAGE and MORE for APA members

A new, medical insurance package for APA members is ready

Five new programs of coverages and deductibles — flexible enough to
meet the differing needs of APA members — offer you and your
colleagues a total medical insurance package.

More than supplementary health-care protection, this package offers.

BASIC MEDICAL — Program A

For the APA member who does not have other medical insurance, Program A provides up to \$5,000 of coverage for each medical condition, subject to a \$50 deductible and an 80/20 co-insurance provision. This provision and the deductible help reduce premium and maintain the low cost group rate available to APA members.

COMPREHENSIVE MEDICAL — Program B

The solution for APA members who desire a combination of Basic and Major Medical Insurance is Program B. While essentially the same as Program A, this Program provides a lifetime maximum benefit of \$250,000.

MAJOR MEDICAL — Programs C, D, E

The programs for APA members who only want to supplement their current health-care plans are the Major Medical Programs. With lower premiums, these offer a range of deductibles — \$10,000, \$15,000, \$25,000 — to suit individual needs.

The medical insurance package you need is ready. And there's more.

HOSPITAL INDEMNITY INSURANCE

With Hospital Indemnity you can safeguard against the additional expenses of hospitalization — the babysitter fees, long distance phone bills, special nursing charges — the expenses that are not covered by your program of health insurance.

Because Hospital Indemnity provides up to \$100 a day, every day that you are hospitalized. A full year of benefits — cash benefits you can use any way you want — with no waiting period before these benefits begin. Hospital Indemnity can help fill in the gaps in your health-care plan.

THE TOTAL MEDICAL INSURANCE PACKAGE AND MORE.

For more information, fill out the coupon below and mail it today.

Send me

- ☐ The Total Medical Insurance Package for APA Members
☐ Hospital Indemnity Insurance

Name _____

Address _____

City _____ State _____ Zip _____

MM 11 9 78



LIBERTY MUTUAL INSURANCE COMPANY
LIBERTY MUTUAL ASSURANCE COMPANY • BOSTON
HOME OFFICE BOSTON

Mail to Mr. Harold Becher, Administrator
APA Insurance Plans
Liberty Mutual Insurance Company
175 Berkeley St.
Boston, Massachusetts 02117

29112
79

Could you use a projective test eliciting a rich variety of responses from various layers of the personality? Easily and quickly administered? Giving results of demonstrated reliability and validity? For children as well as adults?

The
**KAHN TEST
OF SYMBOL
ARRANGE-
MENT**

uses simple plastic objects arranged by the examinee in various sequences according to several different sets of easily followed instructions. An excellent supplement to the TAT and Rorschach.

•

(Obtain information and/or order (complete set \$50.00) from

**PSYCHOLOGICAL TEST
SPECIALISTS**

Box 9229
Missoula, Montana 59807

VOLUME 16, NUMBER 6

DECEMBER 1978

*Library
2/28/79*

JOURNAL OF
CONSULTING
AND
CLINICAL
PSYCHOLOGY

AMERICAN PSYCHOLOGICAL ASSOCIATION

ISSN 0022-006X

- 1368 A and B Undergraduate Interviewers of Schizophrenic and Neurotic Inpatients: A Test of The Interaction Hypothesis/*Daniel F. Barnes and Juris I. Berzins*
- 1374 Drug Abuse Patterns, Personality Characteristics, and Relationships With Sex, Race, and Sensation Seeking/*Patricia B. Suther, Robert P. Archer, and Albert N. Allain*
- 1379 Marital Satisfaction and Depression as Predictors of Physical Health Status/*Robert L. Weiss and Barbara M. Aved*
- 1385 Sexual Fantasies of Females as a Function of Sex Guilt and Experimental Response Cues/*Denise Moreault and Diane R. Follingstad*
- 1388 On the External Validity of Two Psychotherapy Analogues/*Kenneth Kushner*
- 1403 Are Special Norms for Minorities Needed? Development of an MMPI F Scale for Blacks/*Malcolm D. Gynther, David Lachar, and W. Grant Dahlstrom*
- 1409 Parental Personality Factors in Child Abuse/*John J. Spinetta*
- 1415 Effects of Intestinal Bypass Surgery on Body Concept/*Peter M. Silberfarb, Patricia J. Phelps, Peter Hauri, and Charles Solow*
- 1419 Effects of Sex, Social Desirability, and Birth Order on the Defense Mechanisms Inventory/*Gary E. Dudley*
- 1423 Blood Nicotine and Carboxyhemoglobin Levels After Rapid-Smoking Aversion Therapy/*M. A. H. Russell, Martin Raw, C. Taylor, C. Feyerabend, and Y. Saloojee*
- 1432 The Hand Dynamometer as a Neuropsychological Measure/*Carl B. Dodrill*
- 1436 An Initial Look at the Redundancy of Specialized MMPI Scales/*James R. Clopton and Gary L. Klein*
- 1439 Effects of a Self-Control Manual, Rapid Smoking, and Amount of Therapist Contact on Smoking Reduction/*Russell E. Glasgow*
- 1448 A Social-Behavioral Analysis of Skill Deficits in Delinquent and Nondelinquent Adolescent Boys/*Barbara J. Freedman, Lisa Rosenthal, Clyde P. Donahoe, Jr., David G. Schlundt, and Richard M. McFall*
- 1463 Psychosocial Correlates of Empirical Types of Multiple Drug Abusers/*G. Nicholas Braucht, Michael W. Kirby, and G. James Berry*
- 1476 Comparative Evaluation of Therapeutic Components Associated With Behavioral Marital Treatments/*Gayla Margolin and Robert L. Weiss*

Comments

- 1487 Bias or Artifact? A Reply to Reade and Wertheimer/*David J. McDowell*
- 1489 Psychometric Phenology Revisited: Comments on Neuropsychological Testing/*Steven Waksman*
- 1491 Response to Psychometric Phenology Revisited: Comments on Neuropsychological Testing/*Dominic Amante*
- 1493 Can Clients Provide Valuable Feedback to Clinicians About Their Personality Interpretations? A Reply to Greene/*C. R. Snyder, Mitchell M. Handelsman, and Janet R. Endelman*
- 1496 Can Clients Provide Valuable Feedback to Clinicians About Their Personality Interpretations? Greene Replies/*Roger L. Greene*
- 1498 Premature Conclusions Regarding Black and White Suicide Attempters: A Reply to Steele/*Richard C. Bedrosian and Aaron T. Beck*
- 1500 Steele's Reply to Bedrosian and Beck/*Robert E. Steele*
- 1502 Psychotherapy or Massage Parlor Technology? Comments on the Zeiss, Rosen, and Zeiss Treatment Procedure/*Kent G. Bailey*
- 1507 Is Masturbation Still Wrong? Comments on Bailey's Comments/*Nathaniel N. Wagner*
- 1510 Ethical and Professional Issues in Sex Therapy: Comments on Bailey's "Psychotherapy or Massage Parlor Technology?"/*G. Terence Wilson*
- 1515 Comments on Levin et al. and Rosen and Kopel: Internal and External Validity Issues/*Gary M. Farkas*
- 1517 Penile Tumescence as a Measure of Sexual Arousal: A Reply to Farkas/*Saul M. Levin, Salvatore Gambaro, and Lawrence Wolfsohn*
- 1519 Role of Penile Tumescence Measurement in the Behavioral Treatment of Sexual Deviation: Issues of Validity/*Raymond C. Rosen and Steven A. Kopel*
- 1522 A Note on Psychosomatic Factors in the Etiology of Neoplasms/*Jonathan Kellerman*

- 1524 Psychosomatic Etiological Factors in Neoplasms: A Response to Kellerman/*Charles G. Watson and Donald Schuld*
- 1526 The Therapist-as-Fixed-Effect Fallacy in Psychotherapy Research/*Colin Martindale*
- 1531 Resurrection of Uniformity Assumption Myths and the Fallacy of Statistical Absolutes in Psychotherapy Research/*Gordon L. Paul and Mark H. Licht*
- 1535 A Question as to the Validity of the Verbal Scale IQ as a WAIS Short Form/*William M. Reynolds*
- 1537 Validity of the Verbal Scale IQ as a WAIS Short Form: A Reply to Reynolds/*Robert W. Wildman and Robert W. Wildman II*

Brief Reports

- 1538 Effects of Type of Social Reinforcement on the Intelligence Test Performance of Lower-Class Black Children/*Francis Terrell, Jerome Taylor, and Sandra L. Terrell*
- 1540 The Child With Cancer: Patterns of Communication and Denial/*John J. Spinetta and Lorrie J. Maloney*
- 1542 Repression-Sensitization and Health Behavior/*William F. Gayton, John E. Bassett, Joseph Tavormina, and Kenneth L. Ozmon*
- 1545 Relationship of Two Process Measurement Systems for Group Therapy/*John E. Roe and Keith J. Edwards*
- 1547 Development and Validation of a Heterosocial Skills Inventory: The Survey of Heterosexual Interactions for Females/*Carolyn L. Williams and Anthony R. Ciminero*
- 1549 Interpersonal Liking and Self-Disclosure/*Richard Gelman and Hugh McGinley*
- 1552 Reported Stressful Events During Developmental Periods and Their Relation to Locus of Control Orientation in College Students/*Stephen Nowicki, Jr.*
- 1554 An Index of Premorbid Intelligence/*Robert S. Wilson, Gerald Rosenbaum, Gregory Brown, Daniel Rourke, Douglas Whitman, and James Grisell*
- 1556 Relationships Among Marital Assessment Procedures: A Correlational Study/*Gayla Margolin*
- 1559 Effects of Uncertainty About the Behavior of a Phobic Stimulus on Subjects' Fear Reactions/*John Lick, Mark Condiotte, and Thomas Unger*
- 1561 Some Useful Statistics for the Interpretation of the WISC-R/*Terry B. Guthin*
- 1564 Symptom Contamination of the Schedule of Recent Events/*Robert E. Lehman*
- 1566 The Polydrug Assessment Scale: A Psychometric Technique for the Indirect Measurement of Drug Use/*Khalil A. Khavari and Frazier M. Douglass IV*
- 1569 Comparison of Symbolic and Overt Aversion in the Self-Control of Smoking/*Oscar A. Barbarin*
- 1572 Moderators of Racial Differences on the MMPI/*Arthur I. Rosenblatt and David A. Pritchard*
- 1574 Problems Associated With the Typological Measurement of Sex Roles and Androgyny/*Jeffrey A. Kelly, Wyndol Furman, and Veronica Young*
- 1577 Male and Female Treatment Differences: Can They Be Generalized?/*Andrew G. Del Gaudio, Paul J. Carpenter, and Gary R. Morrow*
- 1579 Effect of Psychotomimetics (LSD and Dextroamphetamine) on the Use of Figurative Language During Psychoanalysis/*Michael Natale, Michael Kowitz, Charles Dahlberg, and Joseph Jaffe*
- 1581 Psychometric Correlates of the Mosher Forced Choice Guilt Inventory/*Kevin E. O'Grady and Louis H. Janda*
- 1583 Clinical Application of a Broad-Spectrum Behavioral Approach to Chronic Smokers/*Harry A. Lando and John A. McCullough*
- 1586 Relationships Between Client Self-Perceptions of Self-Consciousness Levels and Therapist Awareness of These Perceptions/*Robert G. Turner and Mae Keyson*
- 1588 Situational Management, Standard Setting, and Self-Reward in a Behavior Modification Weight Loss Program/*Stanley L. Chapman and D. Balfour Jeffrey*
- 1590 Psychotherapeutic Styles: Self-Perceptions of Therapists Varying on A-B Type and Experience/*William B. Goodwin*
- 1191 Notice on Author Alterations

INSTRUCTIONS TO AUTHORS

Style of manuscripts. Manuscripts must be prepared in the style described in the *Publication Manual of the American Psychological Association* (2nd ed.). Instruction on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Authors should submit manuscripts in triplicate and should keep a copy to guard against loss.

Abstracts. Manuscripts of regular articles must be accompanied by an abstract of 100-175 words. Manuscripts of brief reports must be accompanied by an abstract of 75-100 words. All abstracts must be typed on a separate sheet of paper.

Brief Reports. The *Journal of Consulting and Clinical Psychology* will accept Brief Reports of research studies in clinical psychology. The procedure is intended to permit the publication of soundly designed studies of specialized interest or limited importance which cannot now be accepted as regular articles because of lack of space. Several pages in each issue will be devoted to Brief Reports, published in the order of their receipt without respect to the dates of receipt of the regular articles. Most Brief Reports appear in the first or second issue to go to press following their final acceptance.

An author who wishes to submit a Brief Report:

1. Sends the Brief Report, limited to two printed pages and prepared according to the specifications given below.

2. Agrees not to submit the full report to another journal of general circulation.

Specifications. The Brief Report should give a clear, condensed summary of the procedure of the study and as full an account of the results as space permits.

To ensure that a brief report does not exceed two printed pages, follow these instructions for typing: (a) Set typewriter to a 48-space line. (b) Type text. (c) Count all lines except abstracts, title, and by-line, including acknowledgments. If you have exceeded 185 lines, shorten the material. (d) If your brief report barely exceeds 75 lines (one printed page), try to edit to 75 lines exactly. Reports slightly exceeding one page will be edited for length in order to use space economically.

In Brief Reports, headings, tables, and references are avoided, or, if essential, must be counted in the 75 lines. This journal no longer requires an extended report, but, if one is available, each Brief Report must be accompanied by the following footnote:

Requests for reprints and for an extended report of this study should be sent to John Doe (give the author's full name and address).

The footnote should be typed on a separate sheet and *not* counted in the 75-line quota.

Effects of Experimental Manipulation of Self-Disclosure on Group Cohesiveness

Barry J. Kirshner

Jewish Social Service Agency of Metropolitan Washington
Rockville, Maryland

Robert R. Dies and Robert A. Brown

University of Maryland

Self-disclosure in 8-hour experiential groups was systematically controlled by providing detailed audiotaped instructions and illustrations through a series of structured exercises. Two levels of self-disclosure (level of intimacy) were established. Four eight-person heterosexually balanced groups were exposed to encounter group tapes that instructed them to share intimate feelings and experiences. Examples of high self-disclosure and openness were presented to clarify the instructions. In contrast, four comparable groups were conducted by encounter group tapes that furnished only moderate levels of personal disclosure and interpersonal sharing. Groups in both the high and low intimacy conditions received the same set of exercises and differed only in the instructions and accompanying behavioral examples. Results of the study indicate that higher levels of disclosure produced greater group cohesiveness, as hypothesized, on four separate measures of the dependent variable. Findings on three different types of self-report instruments were corroborated by an unobtrusive behavioral measure of cohesiveness.

The effects of group cohesiveness in therapy and encounter group contexts have now been widely documented. Yalom (1975), in his popular text on group psychotherapy, surveyed evidence demonstrating that group cohesiveness is an important determinant of positive therapeutic change and produces many results that are considered to mediate successful therapy outcome. Yalom saw this variable as sufficiently important to list it as 1 of 11 curative factors in group therapy, and to devote an entire chapter to it. Other re-

viewers of the experiential group literature have arrived at similar conclusions regarding the importance of cohesiveness (Goldstein, Heller, & Sechrest, 1966). Clarification of the factors that produce cohesiveness could be enormously useful in designing more helpful group experiences.

Although the consequences of group cohesiveness have been relatively well established, its determinants remain less clear. Goldstein et al. (1966) offered a set of hypothesized variables that ostensibly enhance group cohesiveness, including pregroup expectancies, intergroup competition, temporary inclusion of a "deviant plant," resolution of subgroup differences, and verbal reinforcement; only a few studies have tested these inferences. Other investigators have suggested that group composition (Yalom & Rand, 1966) and leadership style (Yalom, 1975) may be important precursors to cohesiveness. One variable that has particular relevance for the present investigation is the role of self-disclosure.

This article is based on the first author's doctoral dissertation submitted in partial fulfillment of the requirements for the PhD degree at the University of Maryland. Dies and Brown served as co-chairman for this research. Computer support was provided by the University of Maryland Computer Science Center.

The authors wish to thank the other committee members for their contributions: Steven Brauth, Bruce R. Fretz, and Charles Gelso.

Requests for reprints should be sent to Barry J. Kirshner, Jewish Social Service Agency, 6123 Montrose Road, Rockville, Maryland 20852.

Yalom (1975) contends that the group therapist greatly influences the self-disclosure level of group members, and that this in turn contributes to cohesiveness and intermember attraction. The leader, as technical expert, prompts members to share personal material and reinforces self-disclosure through a variety of verbal and nonverbal acts. As a model-setting participant, the leader shares personal feelings, reactions, and experiences, and thereby paves the way for similar risk taking among group members.

There is a small body of research relating self-disclosure to various measures of group cohesiveness and interpersonal attraction. Certner (1973) found that subjects had a greater liking for those from whom they had received more intimate divulgements. Analogously, group members who are higher disclosers early in group life have been found to assume high popularity in the group (Hurley, Note 2). In addition, research by Bean (1972) and Query (1964) indicated that groups high in interpersonal openness were also high in cohesiveness.

Unfortunately, the link between self-disclosure and cohesiveness has occurred mainly within the context of correlational research. One exception is a study by Ribner (1974), who manipulated self-disclosure by means of a pregroup contract. Even though Ribner successfully demonstrated that an initial contract emphasizing self-disclosure did indeed influence subsequent group cohesiveness, his study was limited by the brevity of his group sessions (1 hour), the small variability of topical categories discussed within his groups, and by the analogue nature of his study.

The present investigation was an attempt to extend Ribner's (1974) findings to an actual experiential group setting in which participants had convened for the explicit purpose of improving self-understanding and interpersonal awareness. This study also endeavored to extend the correlational and clinical evidence regarding the presumed association between self-disclosure and cohesiveness by subjecting this relationship to experimental scrutiny.

Method

Subjects

Subjects were recruited from undergraduate classes in psychology and education at the University of Maryland. The students were given a brief description of the research project and were informed that the groups would be 8-hour tape-led interpersonal growth groups designed to facilitate awareness of self and others. Participation was strictly voluntary, without the additional incentive of extra course credit. The final sample consisted of 64 subjects divided into eight groups of 4 males and 4 females each. Half of the groups were assigned to the high intimacy condition and half to the low intimacy condition.

Procedure

All groups met for one 8-hour extended session. Structure was provided by an audiotape similar in format to the *Encountertapes* developed by Berzon (e.g., Berzon, Reisel, & Davis, 1969). The tape presented detailed instructions for the group exercises and furnished illustrations of the behaviors expected in the structured group interactions. The instructions and examples were used to systematically control for levels of intimacy or self-disclosure.

Groups in the high and low intimacy conditions received the same set of exercises and differed only in the instructions and accompanying examples. In the low intimacy condition, instructions and illustrations referred to levels of self-disclosure that were nonpersonal to mildly personal, whereas in the high intimacy condition the self-revelations were moderately to highly personal or private. The sequence of structured activities on the tape was as follows:

Introduction and warm-up exercises (35 minutes). All groups began with a brief introduction explaining the nature of the group. Subjects were informed that after each exercise an unrecorded segment of tape would continue to run, corresponding to the time allotted for the exercise. Following this introduction the group members participated in three warm-up exercises to help them become acquainted with each other and to share some relatively nonthreatening experiences. The structured activities were the name game, trust walk, and negotiated combat (Stevens, 1971).

Top secret (75 minutes). This was the first activity to experimentally manipulate the level of intimacy within the groups. Participants were requested to write an anonymous "secret" about themselves on a slip of paper. Members in the low intimacy condition were asked to furnish secrets or personal information that did not have to be highly revealing. Examples were provided on the tape, for example, "I am a junior and still haven't decided on a major yet." In the high intimacy condition, the secrets were to be private and difficult to share with most people. Examples such as "There were times when I felt so miserable I wondered if life was worth living

anymore" were presented on the tape. Group members were then instructed to place their secrets in the top of a box and then after all the statements had been enclosed and the box was shaken to pull out a secret from the bottom. The box was specially constructed for this task and, unknown to group members, contained a false bottom with 12 pre-selected statements to further insure the proper level of intimacy early in the group. Participants were told that several additional secrets were placed in the box beforehand to further insure anonymity. In this way group members would not get suspicious when their secret was not read. Later questioning by the experimenter indicated that the deception was effective. The deception involved in using contrived secrets initially concerned the investigators. However, the particular secrets used were comparable to those generally obtained with this technique in encounter groups, and similar to those furnished by the present subjects. Furthermore, subjects were not disturbed or resentful when the manipulation was revealed, and both groups reported the experience to be meaningful and personally rewarding. So even though the deception was effective, it did not seriously detract from the richness of the experience. After retrieving a statement, members were instructed to take turns reading them to the group and to consider each one in turn for 5 minutes. After all members had read their secrets, a 20-minute group discussion of their reactions ensued. The tape provided guidelines and an illustration of how to conduct a group discussion. This portion of the tape served to further reinforce the experimental manipulation of intimacy by providing either low or high levels of personal group conversation.

Self-disclosure and cohesiveness measures (15 minutes). All group participants completed two questionnaires at this point in the group. These were a self-disclosure measure, willingness to disclose, and a measure of group cohesiveness (Gruen, 1965). These instruments are described in detail below.

Dyadic interaction (90 minutes). Group members were instructed to choose a partner and then to engage in a 10-minute conversation focused on a topic provided by the tape. Following the conversation, participants were told to choose new partners for another 10-minute discussion on a different topic furnished by the tape. This procedure continued until all group members had met with each other. The topics were selected from material rated as being either high or low in intimacy by Taylor and Altman (1966).

Lunch (30 minutes). Participants were instructed in advance to bring lunch for themselves as well as something to share with the rest of the group.

Problem solving task (30 minutes). This exercise is described by Pfeiffer and Jones (1970) and is used to study sharing of information and leadership behavior in small groups. Briefly, a problem is presented to the group, and each member is supplied with certain pieces of information necessary to solve the problem. The task was included to provide a mechanism for getting the group working together again after lunch.

Self-disclosure exercise (110 minutes). This exercise was similar to top secret, but at this point, the disclosures were no longer anonymous. Groups in the high and low intimacy conditions were again given differential instructions and concrete examples to control for level of intimacy. Then, members took turns reading and commenting on their statements, and when they had finished there was a group discussion based on guidelines furnished by the tape.

Group fantasy (20 minutes). This exercise was presented as a "recreational" activity for the group (Stevens, 1971). There was no attempt to manipulate intimacy.

Self-disclosure and cohesiveness measures (20 minutes). The Willingness to Disclose and Gruen questionnaires were readministered along with two additional cohesiveness measures. The first was an index of cohesiveness developed by Schutz (1958), and the second was the Comfortable Interpersonal Distance Scale (Duke & Nowicki, 1972). These are described below.

Closing and review of the group experience (25 minutes). The final portion of the group experience was used to review what happened and to end with a group hug, in which members gathered close together and silently hugged one another as a group. They were told to terminate the hug whenever they pleased. The group experience was then concluded, and the experimenter met with the participants to debrief them and to provide them with the opportunity to express their reactions to the experience.

Experimental Measures

Self-disclosure was systematically manipulated at several key points in the sequence of structured activities presented on the tape. A check on the effectiveness of the manipulation was essential, since the hypotheses regarding the link between self-disclosure and cohesiveness could not be adequately tested without a clear demonstration that self-disclosure was indeed systematically controlled. Two different measures were used to verify the efficacy of the manipulation.

The first measure was the Willingness to Disclose Questionnaire, which was administered to group participants after the initial manipulation of intimacy in the top secret exercise. The questionnaire was specifically constructed for this investigation and contained 38 statements of various levels of intimacy across a variety of categories. Group members were required to indicate all items that they would be willing to discuss in their experiential group at that time (roughly 2 hours into the session). The items were selected from a pool of statements rated for intimacy by Taylor and Altman (1966). The sum of the intimacy values for the items selected yielded the Willingness to Disclose score.

The second measure of self-disclosure was a modified version of a rating procedure developed by Dies (Note 1) to assess intimacy level of tape-recorded segments of group interaction. Content was judged along a 7-point continuum ranging from impersonal

and event-oriented conversations on the low end of the scale to very private feelings, reactions, and experiences on the high end. In the present study, two segments of the group interactions were tape recorded—the top secret exercise and the self-disclosure exercise—providing portions of early and late group interaction, respectively. Three-minute samples were taken from the middle of the first, second, and last thirds of each segment, yielding three early and three late ratings for each group. These 3-minute samples were coded and randomized before being rated by two trained judges whose scores were averaged to determine the final rating. Interrater reliability was .80 on the 48 samples evaluated in this study.

Several measures of the dependent variable, cohesiveness, were incorporated into this investigation, based on prior research which suggests that cohesiveness is a multifaceted variable capable of being evaluated from a variety of perspectives (e.g., self-report, behavioral). The first was the Gruen (1965) measure, administered along with the Willingness to Disclose Questionnaire after the top secret exercise. This scale contains four group tasks and six product outcomes for each. Participants were instructed to select the outcome that they thought their group would produce if they engaged in the task. The Gruen scale was the only measure of cohesiveness repeated in this study.

Three additional measures of cohesiveness were used, however. A scale described by Schutz (1958) has been shown to produce fairly consistent validity. It is a seven-item Guttman cumulative index to assess favorability of attitudes toward the group and its members. A third measure of cohesiveness, the Comfortable Interpersonal Distance Scale (Duke & Nowicki, 1972), was developed to assess interpersonal closeness. This measure is in the form of a diagram with seven calibrated lines radiating from a central point. Subjects were instructed to imagine themselves at the center of the diagram and to respond to the other group members as if they were moving toward them along the protruding radii. They were to draw a line on the corresponding radius indicating how close they would allow that particular member to advance. The final measure, the group hug, was included to provide an unobtrusive behavioral measure of cohesiveness. This measure was based on research by Dies and Greenberg (1976), which indicated that physical contact was related to here-and-now feelings of interpersonal closeness. The group hug came at the end of the experiential session and was timed; the duration of the hug was considered an index of cohesiveness.

Results

Separate analyses of variance were computed for each of the independent and dependent variables examined in this study. Results of all variables were analyzed using individual subject scores except for the tape

ratings and group hug, which were designed as total group measures.¹

Self-Disclosure (Independent Variable)

The analysis of self-disclosure was essentially a check on the effectiveness of the experimental manipulation. Two separate self-disclosure scores were available. The Willingness to Disclose Questionnaire was administered twice to the group. The second measure was the judges' ratings of self-disclosure from the tapes: Scores from both early (top secret) and late (self-disclosure exercise) group interaction were available. A repeated measures analysis of variance was carried out for the self-disclosure scores. The results summarized in Table 1 indicate that the treatment effects on both variables were highly significant, thus confirming the efficacy of the experimental manipulation.

Group Cohesiveness (Dependent Variable)

Results summarized in Table 1 indicate that higher levels of intimacy or self-disclosure produced greater group cohesiveness on all four dependent measures ($p < .01$). The Gruen instrument was administered at two different times to the group and was therefore analyzed with a repeated measures analysis of variance. The other three measures were administered at the conclusion of the 8-hour marathon. Results of the three self-report measures were confirmed by the unobtrusive behavioral measure of cohesiveness, the group hug.

Development Over Time

Analysis of the Willingness to Disclose instrument indicated that self-disclosure significantly increased over time ($p < .01$), and the tape ratings reflected a trend ($p < .10$) in the same direction. Analysis of the Gruen

¹Group analyses were, in fact, calculated for all variables, since it could be argued that individual scores within groups were not truly independent observations. These group analyses supported the individual analyses in virtually every case. Only the individual analyses are included here to simplify the discussion and presentation of results.

scale, the only measure of cohesiveness repeated in the investigation, also demonstrated a statistically reliable increase as a function of time ($p < .01$). The means and corresponding F values for these analyses appear in Table 1.

The repeated measures design also indicated a significant interaction effect between willingness to disclose and treatment condition ($p < .01$), and between cohesiveness and treatment condition ($p < .01$). Tukey's method for multiple comparisons was computed to test the differences between all means. In terms of willingness to disclose, there was a significant difference between treatment conditions, both early and late in the group. However, the increase over time was much greater for the low intimacy condition groups than for those groups in the high intimacy condition. However, although differences in cohesiveness were significant between high and low conditions at both times in the group life, the greater increase over time on the Gruen measure was found in the high disclosing condition. Apparently, self-disclosure and cohesiveness showed different developmental trends as the groups progressed.

Discussion

Group practitioners agree on the importance of self-disclosure in the formation of meaningful interpersonal relationships. Yalom (1975), for example, theorized that as disclosures proceed in the group, the entire membership gradually increases its level of involvement, responsibility, and obligation to one another: "If the timing is right, there is nothing which will commit an individual to a group more than to receive or to reveal some intimate secret material" (p. 360). Results of the present investigation strongly support this hypothesized relationship between self-disclosure and cohesiveness.

The terms *self-disclosure* and *cohesiveness* have been defined in a variety of ways. With respect to self-disclosure, this study focused on only one of its many salient dimensions, namely, level of intimacy. The influence of other parameters—breadth, frequency, time, and content orientation—were not addressed, nor have they received sufficient attention by

Table 1
Mean Scores (Main Effects) and F Values

Measure	M	$F(1, 62)$
Self-disclosure		
Willingness to disclose		
Treatment (A)		
Low	180.50	58.00*
High	244.20	
Time (B)		
Early	197.27	91.14*
Late	227.42	
A \times B	— ^a	10.22*
Tape rating		
A		
Low	2.00	52.02 ^{b*}
High	4.08	
B		
Early	2.75	5.68*
Late	3.33	
A \times B	— ^a	.12
Cohesiveness		
Gruen (1965)		
A		
Low	11.58	16.45*
High	14.31	
B		
Early	11.21	111.72*
Late	14.69	
A \times B	— ^a	7.30*
Schutz (1958) ^d		
Low	4.70	7.50*
High	5.65	
Comfortable Interpersonal Distance Scale ^d		
Low	33.48	8.46*
High	39.28	
Group hug ^d		
Low	67.50	15.11 ^{b*}
High	140.00	

^a Cell means are available on request.

^b $df = 1, 6$.

^c Trend toward significance ($p < .10$).

^d Only treatment effects are reported, since the measures were only administered once.

* $p < .01$.

other researchers. Yet it is conceivable that these factors are just as influential, or perhaps even more so, than level of intimacy.

Similar arguments could be offered for the concept of "group cohesiveness." Prior researchers have often equated cohesiveness with attraction or liking, whereas the present study defined cohesiveness in several ways: liking, physical contact, comfort with inter-

personal distance, and estimates of group productivity. Although these measures were intercorrelated in this investigation, the moderate level of the correlations suggests that they are not identical scales. A multimethod approach in future studies of group cohesiveness is strongly recommended.

The specific mechanisms accounting for the empirical association between disclosure and cohesiveness have received comparatively scant attention in the literature. Altman and Taylor (1973), however, have suggested that the relationship between self-disclosure and liking (one aspect of cohesiveness) is mutually reciprocal and based on interpersonal rewards and costs. Self-disclosure could be rewarding and thereby enhance liking if it implies that the receiver is trusted (Worthy, Gary, & Kahn, 1969); the more intimate the disclosure, the more rewarding it is for the person receiving it (Certner, 1973). Moreover, disclosures are rewarding if they reveal attitude similarity between the individuals involved (Bersheid & Walster, 1969), or if they imply that the discloser likes the other person (Walster, 1965).

Other possible mechanisms for relating self-disclosure and cohesiveness pertain more to costs than to rewards. For example, highly intimate revelations, rather than being simply rewarding, could make the disclosers feel vulnerable. Cohesiveness may develop as a source of protection for members to insure that their "secrets" are kept within the group. As Yalom (1975) stated,

the receiver . . . is likely to consider himself charged with certain responsibilities or obligations to the discloser. He generally responds to the disclosure by some appropriate comment . . . and then reciprocates with some disclosure of his own. The receiver now, as well as the original discloser, is vulnerable, and a deepening relationship usually continues, with the participants making slightly more open and intimate disclosures in turn until some optimal level for that relationship is reached. (pp. 359-360)

Clarification of the mechanisms involved in the disclosure-cohesiveness relationship is of considerable theoretical and practical importance. For example, group experiences designed to enhance cohesiveness would be dramatically different if the major determinants of cohesiveness were associated with

rewards rather than with interpersonal costs (e.g., attitude similarity as opposed to anxiety over vulnerability).

Even without clarification of these mechanisms, however, group practitioners who wish to enhance group cohesiveness may apparently do so by concentrating on methods of increasing self-disclosure. A variety of procedures can facilitate this process. As a first step, the practitioner can pay careful attention to group composition. Evidence suggests that groups composed of members with high interpersonal skills (including self-disclosure) may be more cohesive (D'Augelli, 1973). The group leader can also introduce a pregroup contract to influence members' expectancies regarding the group experience. A variety of role-induction techniques have been studied including preparatory interviews, written contracts, audiotaped and filmed examples of "appropriate" client behavior, and interpersonal skills practice. A review of this literature indicates that these procedures are effective in enhancing levels of self-disclosure, interpersonal feedback, and other positive member behaviors in both therapy and brief encounter group contexts (Bednar, Melnick, & Kaul, 1974).

Once the group sessions have begun, the clinician can continue to augment member self-disclosure. One method, used in the present study, is to use verbal and nonverbal structured group exercises. Egan (1976), for instance, presented a series of structured activities specifically designed to foster intimate sharing among group participants. Dozens of other compendiums of similar techniques are now available on the market. Some research suggests that these structured techniques can heighten feelings of intermember closeness and willingness to take risks within the group setting (Dies & Greenberg, 1976). The group leader can also increase self-disclosure among group members by adopting a relatively transparent style of leadership (Dies, 1977a), by reinforcing such openness among group participants (Dies, 1977b), and by facilitating the establishment of relatively clear group norms (Lieberman, Yalom, & Miles, 1973).

There were differences between the rates of increase in self-disclosure and cohesiveness.

Even though cohesiveness increased over time, it apparently did not directly follow the growth in willingness to self-disclose. Consequently, one cannot conclude that the increment in readiness to self-disclose over time produced the increased cohesiveness. It is possible that the production of high levels of intimacy early in the group, as opposed to a gradual development over time, was the crucial factor leading to heightened cohesiveness. This interpretation received some earlier support from Hurley's (Note 2) findings that individuals high in disclosing behavior early in their group achieved high popularity later in their group. However, this issue is not entirely clear in the present study, since there were possibly ceiling effects on the willingness to disclose measure.

The form of the relationship between self-disclosure and cohesiveness and the question of which comes first pose important theoretical and pragmatic questions. Nevertheless, this study suggests that we can systematically enhance the level of these variables within our groups and thereby increase the probability of positive therapeutic outcomes.

Reference Notes

1. Dies, R. R. *Self-disclosure in group interactions*. Unpublished manuscript, University of Maryland, 1972.
2. Hurley, J. *Self-disclosure in small counseling groups*. Unpublished manuscript, Michigan State University, 1967.

References

- Altman, I., & Taylor, D. A. *Social penetration: The development of interpersonal relationships*. New York: Holt, Rinehart & Winston, 1973.
- Bean, H. B. The effects of a role-model and instructions on group interpersonal openness and cohesiveness (Doctoral dissertation, West Virginia University, 1972). *Dissertation Abstracts International*, 1972, 37, 4201B. (University Microfilms No. 72-5152, 75)
- Bednar, R. L., Melnick, J., & Kaul, T. J. Risk, responsibility and structure: A conceptual framework for initiating group counseling and psychotherapy. *Journal of Counseling Psychology*, 1974, 21, 31-37.
- Bersheid, E., & Walster, E. *Interpersonal attraction*. Reading, Mass.: Addison-Wesley, 1969.
- Berzon, B., Reisel, J., & Davis, D. P. PEER: An audiotape program for self-directed small groups. *Journal of Humanistic Psychology*, 1969, 9, 71-86.
- Certner, B. C. Exchange of self-disclosures in same-sexed groups of strangers. *Journal of Consulting and Clinical Psychology*, 1973, 40, 292-297.
- D'Augelli, A. Group composition using interpersonal skills: An analogue study on the effects of members' interpersonal skills on peer ratings and group cohesiveness. *Journal of Counseling Psychology*, 1973, 20, 531-534.
- Dies, R. R. Group therapist transparency: A critique of theory and research. *International Journal of Group Psychotherapy*, 1977, 27, 177-200 (a).
- Dies, R. R. Pragmatics of leadership in psychotherapy and encounter group research. *Small Group Behavior*, 1977, 8, 229-248 (b).
- Dies, R. R., & Greenberg, B. Effects of physical contact in an encounter group context. *Journal of Consulting and Clinical Psychology*, 1976, 44, 400-405.
- Duke, M. P., & Nowicki, S. A new measure and social learning model for interpersonal distance. *Journal of Experimental Research in Personality*, 1972, 6, 119-132.
- Egan, G. *Interpersonal living: A skills/contract approach to human-relations training in groups*. Monterey, Calif.: Brooks/Cole, 1976.
- Goldstein, A. P., Heller, K., & Sechrest, L. B. *Psychotherapy and the psychology of behavior change*. New York: Wiley, 1966.
- Gruen, W. A contribution toward understanding of cohesiveness in small groups. *Psychological Reports*, 1965, 17, 311-322.
- Lieberman, M. A., Yalom, I. D., & Miles, M. B. *Encounter groups: First facts*. New York: Basic Books, 1973.
- Pfeiffer, J. W., & Jones, J. E. *A handbook of structured experiences for human relations training*. Iowa City, Iowa: University Associates Press, 1970.
- Query, W. T. Self-disclosure as a variable in group psychotherapy. *International Journal of Group Psychotherapy*, 1964, 14, 107-115.
- Ribner, N. G. Effects of an explicit group contract on self-disclosure and group cohesiveness. *Journal of Counseling Psychology*, 1974, 21, 116-120.
- Schutz, W. C. *FIRO: Three-dimensional theory of interpersonal behavior*. New York: Holt, Rinehart & Winston, 1958.
- Stevens, J. O. *Awareness: Exploring, experimenting, experiencing*. Moab, Utah: Real People Press, 1971.
- Taylor, D. A., & Altman, I. Intimacy-scaled stimuli for use in studies of interpersonal relations. *Psychological Reports*, 1966, 19, 729-730.
- Walster, E. The effect of self-esteem on romantic liking. *Journal of Experimental Social Psychology*, 1965, 1, 184-197.
- Worthy, M., Gary, A. L., & Kahn, G. M. Self-disclosure as an exchange process. *Journal of Personality and Social Psychology*, 1969, 13, 59-63.
- Yalom, I. D. *The theory and practice of group psychotherapy*. New York: Basic Books, 1975.
- Yalom, I. D., & Rand, K. Compatibility and cohesiveness in therapy groups. *Archives of General Psychiatry*, 1966, 13, 267-276.

Subtle-Obvious Ratings of MMPI Items: New Interest in an Old Concept

William L. Christian, Barry R. Burkhart, and Malcolm D. Gynther
Auburn University

Previous attempts to devise subtle-obvious ratings for Minnesota Multiphasic Personality Inventory (MMPI) items have been too gross, restricted to certain scales, or have failed to consider the perspective of the psychologically naive inventory user. To correct these deficiencies, university students were asked to rate all 566 MMPI items, answered true and false, on a 5-point subtle-obvious scale. Reliability of ratings was assessed via responses to duplicate items; the reliability coefficients were .98 for items answered true and .91 for items answered false. Males and females gave very similar ratings; correlations were .94 and .90 for true and false items, respectively. Mean ratings for all subjects showed that MMPI scales *F* and *Sc* were most obvious and that *Mf* and *Si* were least obvious with regard to pathology. The correlation between obviousness ratings and desirability ratings was $-.78$. Analysis of previous measures of the subtlety-obviousness dimension in terms of the present item ratings indicate that previous measures did not adequately represent the very subtle or very obvious extremes of the distribution. Further research with these ratings should demonstrate whether empirically derived inventories with their subtle items are, in fact, useful tools for personality assessment.

Responses to personality inventories are determined by instructional sets, stylistic features, and certain other characteristics of item content, that is, social desirability, sentence structure, ambiguity, and subtlety. Content of Minnesota Multiphasic Personality Inventory (MMPI) items was considered relatively unimportant for many years, partly as a function of Meehl's (1945) influential article. However, the pendulum has swung the other way; more and more attention is now being paid to content (Jackson, 1971; Koss & Butcher, 1973; Wiggins, 1966).

Measures of desirability have long been available (Edwards, 1957; Heineman, 1960; Messick & Jackson, 1961) for those who wish to explore the relations between MMPI item endorsement and desirability. Sentence structure of MMPI items has been analyzed (Wiggins, 1964), and ambiguity ratings are available (Harris & Baxter, 1965). (All of these

"itemmetric" data are available in Dahlstrom, Welsh, & Dahlstrom, 1975.) However, subtle-obvious ratings (despite appearances to the contrary) have not been developed. This lack is curious when one considers that the subtlety-obviousness distinction has been an integral part of the controversy over empirical versus intuitive internal test construction strategies.

There have been three efforts to deal with the subtle-obvious issue: X and O classification (Meehl & Hathaway, 1946), Wiener's (1948) subtle-obvious keys, and Duff's (1965) subtle-intermediate-obvious categories. X items are those that relatively few normals endorsed in the scored direction. Content of these items is usually undisguised and obviously pathological. O items are those endorsed by a majority of normal subjects, but they are endorsed by an even higher percentage of clinical patients (and consequently scored in the direction of pathology). Content of these items is heterogeneous and not especially indicative of emotional disturbance. Wiener and Harmon (Note 1), using a rational-intuitive

Requests for reprints should be sent to Barry R. Burkhart, Department of Psychology, Auburn University, Auburn, Alabama 36830.

procedure, divided all MMPI items into those to which endorsement clearly indicated emotional disturbance and those to which emotional disturbance was not indicated by endorsement. Subtle and obvious keys were actually developed only for the *D*, *Hy*, *Pd*, *Pa*, and *Ma* scales (cf. Dahlstrom, Welsh, & Dahlstrom, 1972). Duff (1965) asked graduate students to identify the clinical scale from which an item was drawn and its keyed response. He used the 162 items from the *Hy*, *Pd*, and *Sc* scales plus 34 items from other clinical scales and 30 items not included on any clinical scale. Items whose scales and pathological responses were identified by no more than 10% of the judges were assigned to the subtle category, items identified by 11%–50% of the judges were assigned to the intermediate category, and items correctly placed by at least 50% of the judges were assigned to the obvious category.

Cronbach (1970), in discussing the MMPI, stated that "separate scoring of subtle and transparent items is emphatically needed" (p. 532). He pointed out that if subtle and obvious keys for the different scales were available, it would be possible to determine where the discriminating power lies. If the subtle keys proved to be invalid or the obvious items carried most of the discriminating power, as suggested by the work of Duff (1965) and Koss and Butcher (1973), the whole criterion-keyed empirical approach to test construction would be called into question. From the point of view of psychometric theory, the most important aspect of this research is the provision of data whereby these issues can be investigated in a more comprehensive manner than previously possible.

No investigator has used standard scaling procedures to determine mean subtle-obvious ratings of all MMPI items. In this study, such ratings were obtained from individuals relatively naive with respect to psychological inventories. Although college students, as a group, tend to be more defensive than men and women in general and, consequently, are more likely to perceive items as containing pathological content, they are more representative of those to whom the inventory is

typically administered than are graduate students or professional psychologists.

Method

Subjects

One hundred thirty-eight undergraduate students enrolled in introductory psychology sections served as subjects. The subjects were divided into three groups, each containing 25 males and 21 females.

Procedure

The first group of subjects rated 387 MMPI items, namely, Items 1–283, answered true, and the first 104 of the 209 items keyed false on the clinical scales answered false. The second group rated Items 284–566 answered true and the 105 remaining items keyed false on the clinical scales answered false. The third group of subjects rated the other 357 items answered false. Subjects were instructed to read each item carefully and to decide how clearly each item was indicative of a psychological problem. Very obvious items were to be assigned a rating of 5; obvious, a rating of 4; neither obvious nor subtle, a rating of 3; subtle, a rating of 2; and very subtle, a rating of 1.

Mean ratings of each item by males, females, and males and females combined were computed. Mean obviousness scores for standard MMPI scales were determined. Relationships between male–female ratings and desirability–obviousness ratings were evaluated by correlational analyses. The mean obviousness values of the X and O items; Wiener's (1948) subtle and obvious items; and Duff's (1965) subtle, intermediate, and obvious items, were calculated using our mean item ratings. Percentages of items in the MMPI clinical scales that fell into each of the five rating categories also were calculated.

Results and Discussion

Analyses of the ratings of the duplicate items in the MMPI showed very high agreement between the groups of judges. The product-moment correlation for the duplicate items answered true was .98 ($n = 14$) and for the duplicate items answered false was .91 ($n = 5$).¹ Means and standard deviations for the subtle-obvious ratings of all MMPI items answered true and false are contained in

(text continued on p. 1184)

¹ Although there are 16 duplicate items in the MMPI, the reliability coefficients are based on less than the full 16 items, because in dividing the items for rating by the groups of judges, some of the duplicate items were not rated by two groups of judges.

Table 1
Means and Standard Deviations of Obviousness Ratings for Minnesota Multiphasic Personality Inventory Items

True					False					Item no.	True					False					Item no.	True					False				
M	SD	M	SD	M	SD	M	SD	M	SD		M	SD	M	SD	M	SD	M	SD	M	SD		M	SD	M	SD	M	SD				
1	1.11	.53	1.44	.86	37	1.87	.95	4.11	1.04	73	3.02	1.24	3.13	1.31	109	3.04	1.03	2.44	1.15												
2	1.22	.55	2.91	1.15	38	2.89	1.04	2.17	1.17	74	3.11	1.66	2.71	1.46	110	4.28	1.03	2.40	1.38												
3	1.37	.61	3.00	1.21	39	3.89	.95	2.33	1.06	75	1.76	1.04	3.24	1.25	111	1.96	1.12	3.09	1.30												
4	1.83	.97	1.60	1.05	40	3.78	1.01	2.38	1.11	76	4.00	.82	2.06	1.30	112	1.63	.95	3.33	1.10												
5	2.37	.95	1.88	1.15	41	4.11	.99	2.48	1.18	77	1.67	.79	2.46	1.21	113	1.39	.83	3.98	1.02												
6	3.00	1.03	2.35	1.06	42	2.65	1.20	2.06	1.14	78	1.39	.68	1.88	1.00	114	3.52	1.15	2.17	1.22												
7	1.83	1.02	2.67	1.25	43	4.17	.83	2.27	1.37	79	1.89	1.10	3.28	1.11	115	1.70	1.19	2.87	1.31												
8	1.48	.86	3.46	1.11	44	3.94	1.00	2.15	1.27	80	3.09	1.09	2.17	1.24	116	2.96	1.01	2.15	1.23												
9	2.04	1.05	2.91	1.09	45	3.28	1.30	2.17	1.24	81	1.35	.80	1.59	.88	117	2.80	1.28	2.46	1.30												
10	3.41	1.07	2.02	1.26	46	1.94	.88	3.33	1.01	82	2.91	.84	2.11	1.08	118	2.41	1.13	2.15	1.14												
11	3.28	1.03	2.58	1.21	47	3.52	1.03	2.13	1.16	83	1.30	.70	3.49	1.36	119	1.89	1.16	3.04	1.30												
12	1.52	.81	2.09	1.23	48	4.09	1.17	2.54	1.29	84	3.07	1.37	2.81	1.31	120	1.76	1.04	2.22	1.19												
13	3.22	1.15	2.21	1.02	49	4.07	.98	2.12	1.48	85	4.61	.75	2.25	1.48	121	4.41	.93	2.13	1.37												
14	3.22	1.21	2.35	1.40	50	4.76	.71	2.27	1.43	86	3.44	1.03	2.38	1.32	122	1.67	.94	3.67	1.01												
15	3.54	1.21	2.74	1.31	51	1.52	.86	2.76	1.06	87	1.39	.83	1.71	.91	123	4.22	.96	2.08	1.28												
16	4.35	1.13	2.25	1.37	52	3.11	1.04	2.17	1.20	88	1.65	.70	4.26	1.10	124	3.02	1.13	2.50	1.03												
17	1.50	.85	3.33	1.38	53	3.70	1.25	2.13	1.19	89	2.63	.95	2.56	1.17	125	3.11	1.18	2.06	1.30												
18	1.76	1.12	2.54	1.30	54	2.02	1.16	4.24	.97	90	1.96	1.07	2.00	1.15	126	1.61	.98	2.02	1.18												
19	3.13	1.02	2.39	1.06	55	1.74	1.10	2.96	1.25	91	2.74	1.34	2.65	1.25	127	3.35	1.29	2.75	1.28												
20	1.87	1.09	3.35	1.23	56	2.70	1.20	2.02	.98	92	1.52	1.07	1.85	1.04	128	1.78	1.01	2.65	1.23												
21	3.46	1.19	2.23	1.13	57	1.65	.77	3.15	1.17	93	3.25	1.12	2.59	1.22	129	2.89	.99	2.30	1.13												
22	4.20	1.13	2.48	1.35	58	2.50	1.17	2.80	1.20	94	3.65	1.10	2.48	1.26	130	1.65	1.12	3.13	1.19												
23	3.89	1.18	2.00	1.22	59	2.48	1.09	2.31	1.18	95	1.52	1.01	2.20	1.16	131	1.85	1.01	3.44	1.26												
24	4.24	.99	2.50	1.36	60	1.44	.91	1.80	1.05	96	1.35	.67	3.35	1.27	132	1.41	.83	2.06	1.27												
25	1.50	.98	1.52	.89	61	3.22	1.21	2.33	1.26	97	4.15	1.01	2.38	1.16	133	1.87	1.15	3.35	1.32												
26	2.54	1.13	2.46	1.21	62	4.00	.97	2.52	1.31	98	1.63	1.12	2.67	1.52	134	1.76	.90	2.67	1.23												
27	4.76	.71	2.31	1.57	63	1.72	1.19	3.26	1.26	99	1.89	1.04	2.98	1.25	135	2.87	1.15	1.80	1.00												
28	4.02	1.00	1.78	1.05	64	3.22	1.05	2.07	1.14	100	2.59	1.12	2.75	1.10	136	3.50	1.24	2.46	1.35												
29	3.30	1.15	1.92	1.12	65	1.53	1.01	4.17	1.12	101	1.63	1.02	3.29	1.29	137	1.59	.91	3.48	1.13												
30	2.46	1.09	2.44	1.24	66	4.78	.73	2.33	1.52	102	2.76	1.21	2.75	1.28	138	2.98	1.22	3.00	1.14												
31	3.67	1.05	2.23	1.23	67	3.65	1.10	3.46	1.21	103	1.65	.99	3.46	1.05	139	4.63	.74	2.27	1.43												
32	3.50	1.00	2.12	1.04	68	1.89	1.12	3.00	1.16	104	4.33	.70	2.21	1.47	140	1.28	.72	1.98	1.08												
33	3.54	1.05	2.44	1.33	69	4.15	1.10	2.54	1.72	105	2.11	.99	2.59	1.29	141	2.41	1.00	2.80	1.26												
34	2.94	1.25	2.10	1.26	70	1.96	1.15	2.38	1.22	106	4.35	.85	2.52	1.42	142	3.24	1.06	2.07	1.00												
35	4.54	.69	2.58	1.35	71	2.61	1.20	2.59	1.22	107	1.57	.94	4.02	1.15	143	1.96	1.07	2.52	1.13												
36	1.80	.86	3.33	1.32	72	3.26	1.04	1.87	1.14	108	3.13	1.29	2.15	1.06	144	1.80	1.09	1.94	1.04												

Table 1—(continued)

Item no.	True		False		Item no.	True		False		Item no.	True		False		Item no.	True		False	
	M	SD	M	SD		M	SD	M	SD		M	SD	M	SD		M	SD	M	SD
145	3.74	.98	1.98	1.04	181	2.28	1.05	2.79	1.38	217	3.50	1.13	1.59	1.39	253	2.67	1.32	2.54	1.20
146	3.11	.95	2.15	1.13	182	4.33	.97	2.42	1.41	218	4.13	.83	2.52	1.05	254	2.87	1.20	1.70	.89
147	3.09	.92	1.96	.97	183	2.26	1.04	1.61	.75	219	1.28	.72	1.63	.83	255	2.50	1.21	1.33	.67
148	2.94	1.04	2.04	1.07	184	4.61	.77	2.54	1.57	220	1.35	.80	4.02	1.00	256	2.46	1.05	1.85	1.19
149	1.54	.96	1.88	1.10	185	1.39	.61	2.46	1.17	221	1.28	.81	1.76	1.02	257	1.61	.83	4.04	.97
150	1.59	1.05	3.89	1.10	186	3.57	1.07	2.00	1.12	222	2.50	1.19	2.33	.98	258	1.39	1.09	2.85	1.49
151	4.52	.79	2.19	1.55	187	1.70	1.09	3.35	1.12	223	2.04	1.32	1.61	.83	259	2.72	.91	2.00	1.03
152	1.64	.88	3.50	1.17	188	1.28	.78	2.44	1.33	224	2.44	1.22	1.96	.97	260	2.72	1.03	1.33	.67
153	1.36	.65	3.59	1.07	189	3.35	1.16	1.87	1.10	225	2.02	.86	2.13	1.19	261	1.52	.81	1.94	1.24
154	1.44	.92	3.70	1.15	190	1.46	.84	3.44	1.00	226	2.72	1.22	2.13	1.17	262	1.80	.91	3.15	1.07
155	1.31	.73	3.24	1.16	191	2.70	1.25	1.80	.96	227	3.07	1.25	1.87	.95	263	2.72	1.05	1.46	.78
156	4.29	.94	2.42	1.53	192	1.57	1.11	3.37	1.02	228	1.72	.96	2.77	1.18	264	2.44	1.26	3.04	1.07
157	3.54	1.09	2.13	1.14	193	1.30	.63	3.27	1.23	229	1.80	.96	2.30	1.19	265	3.87	1.00	1.76	.85
158	3.24	.99	2.08	1.06	194	4.26	1.02	2.27	1.50	230	1.54	1.05	3.11	1.16	266	2.80	1.28	2.92	1.49
159	3.04	.94	2.13	1.16	195	2.00	1.16	2.46	1.31	231	2.94	1.16	2.49	1.24	267	2.80	.96	1.35	.77
160	1.35	.74	3.22	1.21	196	1.44	.78	3.00	1.19	232	2.35	1.18	2.35	1.03	268	2.00	1.94	3.07	1.08
161	1.28	.54	2.02	1.15	197	3.48	1.36	2.23	1.25	233	3.41	1.31	1.78	.88	269	3.91	1.05	2.12	1.32
162	2.78	1.03	2.50	1.26	198	1.85	.84	3.26	.98	234	2.87	1.15	1.89	.96	270	2.87	1.29	1.98	1.00
163	1.48	.78	2.85	1.05	199	1.96	1.15	2.91	1.19	235	1.96	.94	2.60	1.07	271	3.67	.99	1.61	.75
164	1.20	.62	3.15	1.17	200	4.20	.96	2.06	1.27	236	3.58	.97	1.92	1.01	272	1.67	1.01	2.59	1.26
165	2.94	1.08	2.83	1.27	201	2.22	1.07	3.22	1.10	237	2.83	1.29	2.49	1.20	273	3.37	1.25	1.85	1.07
166	3.22	1.20	1.91	1.13	202	4.54	.75	2.10	1.51	238	3.59	1.15	2.06	1.02	274	1.52	.89	2.04	1.13
167	3.28	1.19	2.44	1.43	203	1.41	.78	1.63	.82	239	2.54	1.31	2.54	1.34	275	4.76	.60	2.23	1.52
168	4.54	.84	2.52	1.46	204	1.26	.58	1.67	.98	240	2.46	.98	2.35	1.30	276	1.24	.64	3.09	1.31
169	1.61	1.02	3.74	.98	205	4.52	.91	2.29	1.29	241	2.98	1.29	1.83	.97	277	2.96	1.32	2.13	1.12
170	2.28	.98	2.91	1.21	206	2.20	1.31	2.13	1.21	242	1.94	1.02	3.35	1.16	278	3.59	1.13	2.50	1.24
171	2.48	1.09	2.24	1.18	207	1.22	.66	3.17	1.02	243	1.57	.83	3.54	1.05	279	2.15	.99	1.50	.81
172	2.74	1.14	2.00	.97	208	2.24	.95	1.85	.84	244	3.07	1.16	2.21	1.13	280	3.13	1.13	1.74	.93
173	1.37	.77	3.28	1.03	209	4.20	.93	2.04	1.28	245	3.61	1.11	2.12	1.25	281	1.98	1.13	3.78	.94
174	1.37	.88	3.09	1.05	210	3.57	1.17	1.88	1.25	246	2.83	1.31	1.75	1.14	282	3.26	1.31	2.21	1.30
175	1.52	.91	3.44	.98	211	3.78	.96	2.02	1.18	247	3.48	1.03	2.12	1.34	283	1.33	.76	1.67	.92
176	1.63	1.00	3.33	1.35	212	3.48	1.17	2.00	1.07	248	3.20	1.38	2.20	1.20	284	4.09	.90	2.17	1.15
177	1.39	.77	3.76	1.10	213	3.83	1.08	1.41	0.83	249	1.91	1.15	2.27	1.21	285	1.89	1.08	2.44	1.21
178	1.57	.81	3.76	1.12	214	1.54	.89	2.52	1.17	250	2.89	1.22	2.50	1.32	286	3.78	.92	1.81	.99
179	3.26	1.10	2.38	1.07	215	4.22	1.03	2.04	1.39	251	4.26	1.18	2.12	1.29	287	2.33	1.02	3.48	1.07
180	2.61	.93	1.80	1.15	216	3.44	1.17	1.73	1.12	252	4.04	.99	2.13	1.31	288	3.44	1.05	1.88	1.18

Table 1—(continued)

Item no.	True			False			Item no.	True			False			Item no.	True			False			
	M	SD		M	SD			M	SD		M	SD			M	SD		M	SD		
289	1.96	1.10	2.80	1.19	3.24	1.04	325	3.24	1.04	2.23	1.18	361	2.96	1.00	2.10	1.07	397	1.38	.86	2.87	1.02
290	3.17	1.16	2.08	1.15	4.15	1.05	326	4.15	1.05	2.29	1.40	362	3.11	1.05	2.08	1.12	398	3.58	.89	1.59	1.02
291	4.13	1.13	2.06	1.33	3.27	2.50	327	2.50	1.15	2.07	1.02	363	4.36	.93	2.23	1.54	399	3.16	1.04	2.79	1.36
292	2.98	1.06	1.44	.69	3.28	3.48	328	3.48	.98	1.96	1.01	364	3.83	1.04	2.10	1.18	400	2.58	1.37	2.83	1.27
293	4.30	.92	2.23	1.38	3.29	3.41	329	3.41	1.19	1.95	1.19	365	4.00	.94	1.87	1.09	401	2.02	1.20	2.65	1.31
294	1.48	.81	3.15	1.14	3.30	1.74	330	1.74	.98	3.13	1.29	366	4.22	.70	1.94	1.11	402	2.13	.89	2.12	1.04
295	1.74	1.00	2.00	1.07	3.31	4.28	331	4.28	.83	2.27	1.37	367	1.96	1.01	2.25	1.22	403	1.67	.88	3.00	1.14
296	1.98	1.16	2.52	1.07	3.32	3.11	332	3.11	1.10	1.83	1.14	368	2.37	1.04	2.50	1.16	404	2.61	1.15	2.21	1.18
297	3.28	1.26	2.88	1.31	3.33	4.24	333	4.24	.74	2.17	1.29	369	3.30	.87	2.56	1.26	405	1.57	.89	2.71	1.27
298	3.20	1.09	2.08	1.19	3.34	3.37	334	3.37	1.16	1.98	1.11	370	3.02	.95	2.50	1.34	406	3.04	1.19	2.17	1.32
299	2.80	.93	2.13	1.09	3.35	3.64	335	3.64	1.12	2.04	1.19	371	1.54	.94	2.04	1.12	407	1.59	.88	3.12	1.44
300	2.85	1.14	1.76	.92	3.36	3.44	336	3.44	.87	2.06	1.09	372	2.89	.95	1.75	1.03	408	2.94	1.20	2.38	1.09
301	4.13	.96	2.15	1.16	3.37	4.17	337	4.17	.85	2.02	1.11	373	3.22	1.01	2.38	1.29	409	2.37	1.06	2.21	1.09
302	1.72	1.09	3.54	1.10	3.38	3.87	338	3.87	.89	2.42	1.24	374	2.24	1.08	3.02	1.24	410	2.65	1.14	2.77	1.20
303	3.78	.87	2.17	1.22	3.39	4.78	339	4.78	.76	2.17	1.42	375	2.48	1.03	2.29	1.11	411	2.18	1.19	2.38	1.29
304	3.13	.98	1.94	1.19	3.40	2.57	340	2.57	1.17	2.27	1.25	376	3.59	.96	2.90	1.24	412	1.83	1.04	2.38	1.07
305	4.09	.69	2.04	1.19	3.41	3.39	341	3.39	1.18	1.96	1.03	377	2.33	1.16	1.71	.96	413	4.00	1.03	2.62	1.39
306	2.48	1.13	3.63	.97	3.42	3.94	342	3.94	1.02	1.98	1.11	378	2.46	1.09	2.25	1.20	414	3.89	.91	2.02	1.04
307	2.80	.98	1.94	1.07	3.43	3.20	343	3.20	1.13	2.15	1.14	379	2.20	.98	2.79	1.19	415	2.00	.95	2.27	1.14
308	3.20	1.13	3.15	1.13	3.44	3.44	344	3.44	1.00	2.13	1.17	380	1.57	.89	2.46	1.11	416	2.76	1.07	2.17	1.20
309	1.72	.96	1.39	.91	3.45	3.89	345	3.89	1.11	2.15	1.14	381	2.24	.82	1.77	0.94	417	2.89	1.19	2.25	1.15
310	1.72	.86	3.46	1.01	3.46	3.09	346	3.09	1.17	1.98	1.18	382	1.54	.81	3.15	1.06	418	4.22	.77	2.37	1.28
311	2.80	1.17	1.87	1.10	3.47	2.00	347	2.00	1.01	3.83	1.12	383	1.65	.95	2.65	1.06	419	2.78	1.09	1.81	1.01
312	4.11	1.06	1.90	1.16	3.48	3.00	348	3.00	1.17	1.85	1.03	384	1.94	1.08	2.25	1.30	420	3.07	1.12	2.15	1.11
313	2.78	1.11	2.24	1.18	3.49	3.57	349	3.57	1.11	2.40	1.29	385	2.50	1.07	2.27	1.22	421	2.83	1.04	1.90	1.09
314	2.98	1.16	1.91	1.03	3.50	4.04	350	4.04	.92	2.25	1.14	386	2.24	.99	2.00	.97	422	3.33	1.01	1.81	1.01
315	4.31	.90	2.19	1.36	3.51	3.76	351	3.76	.92	1.87	1.05	387	1.89	1.14	2.17	1.28	423	1.49	.82	1.77	.96
316	3.13	.92	1.96	1.05	3.52	3.89	352	3.89	.96	2.23	1.28	388	2.89	1.11	1.90	1.03	424	3.17	1.20	1.92	1.10
317	2.98	1.01	2.02	.92	3.53	1.76	353	1.76	1.17	3.45	1.11	389	3.48	1.05	2.00	1.05	425	1.98	1.00	2.38	1.29
318	1.62	1.01	3.28	.78	3.54	3.80	354	3.80	1.03	2.06	1.24	390	1.46	.72	2.69	1.08	426	2.57	1.15	2.27	1.10
319	3.33	1.02	1.74	.95	3.55	4.64	355	4.64	.88	2.10	1.49	391	2.30	.96	1.76	1.08	427	2.00	.89	2.29	1.13
320	2.98	1.22	2.08	1.03	3.56	3.61	356	3.61	.88	1.88	.98	392	1.76	.95	2.12	1.15	428	1.54	.75	1.87	1.03
321	2.71	1.01	1.85	1.04	3.57	3.49	357	3.49	.99	2.19	1.21	393	3.67	.79	2.48	1.48	429	1.67	.76	1.77	1.00
322	2.63	.90	1.78	.91	3.58	3.83	358	3.83	.97	2.31	1.32	394	1.94	.98	2.40	1.14	430	1.33	.67	3.96	1.51
323	3.31	1.24	1.72	1.00	3.59	3.33	359	3.33	1.04	1.74	.88	395	3.15	1.25	2.17	1.25	431	3.41	.98	2.54	1.21
324	3.70	1.17	1.94	1.21	3.60	3.98	360	3.98	1.03	2.08	1.22	396	1.72	.91	2.15	1.30	432	1.76	.92	2.00	.95

Table 1—(continued)

Item no.	True		False		Item no.	True		False		Item no.	True		False		Item no.	True		False	
	M	SD	M	SD		M	SD	M	SD		M	SD	M	SD		M	SD	M	SD
433	3.67	1.35	2.08	1.33	467	2.72	1.24	2.12	1.17	500	2.67	1.10	2.31	1.11	534	2.33	1.08	2.15	.96
434	1.65	.90	1.67	1.00	468	2.54	1.24	2.73	1.21	501	2.02	1.00	2.29	1.07	535	3.26	1.16	1.83	.96
435	2.33	1.02	2.25	1.19	469	2.89	1.20	1.52	.96	502	3.07	1.22	1.52	.94	536	2.48	.96	2.25	1.14
436	2.64	1.26	2.42	1.14	470	4.28	.72	2.27	1.37	503	2.59	1.13	2.37	1.22	537	2.17	1.16	1.88	1.04
437	3.00	1.11	2.19	1.12	471	3.28	.91	2.04	1.12	504	2.71	.94	2.19	1.03	538	1.87	1.05	1.81	1.03
438	3.33	1.18	2.23	1.18	472	3.65	1.14	2.06	1.09	505	4.30	.99	1.89	1.35	539	1.87	1.00	2.23	1.23
439	3.00	1.19	1.98	1.09	473	3.17	1.14	2.10	1.01	506	3.35	1.08	1.98	1.08	540	1.83	1.12	2.50	1.42
440	2.96	1.08	1.83	1.04	474	1.80	.98	2.13	1.03	507	3.02	1.04	1.96	1.08	541	2.83	1.12	2.02	1.15
441	2.11	.97	1.90	.98	475	2.96	1.13	2.33	1.10	508	1.72	.89	2.02	1.02	542	2.22	1.21	2.31	1.23
442	2.86	1.27	2.46	1.20	476	3.85	1.21	2.27	1.51	509	2.91	1.13	2.13	1.05	543	4.02	.98	2.31	1.38
443	3.34	.91	2.42	1.21	477	2.85	1.15	2.10	1.19	510	3.98	1.09	2.17	1.20	544	3.35	1.02	1.83	.94
444	2.15	1.01	2.42	1.19	478	2.89	1.04	2.33	1.17	511	3.72	1.09	2.25	1.30	545	3.15	1.01	1.98	1.08
445	1.76	.90	3.00	1.47	479	2.87	1.13	1.91	1.13	512	4.09	.87	2.02	1.32	546	1.54	.89	1.90	.98
446	1.89	.90	3.00	1.23	480	3.59	.88	1.98	1.06	513	2.22	1.13	2.08	1.01	547	3.15	1.17	1.78	1.05
447	2.98	1.04	2.25	1.23	481	3.61	1.04	1.57	1.05	514	3.65	1.23	2.21	1.18	548	2.74	1.24	2.60	1.30
448	3.96	.94	2.27	1.32	482	3.44	.96	1.57	.91	515	1.72	.98	2.21	1.24	549	3.72	1.00	2.35	1.15
449	2.59	1.17	2.22	1.09	483	1.67	.87	2.87	1.47	516	2.37	.97	1.88	.88	550	2.00	1.16	1.77	.94
450	1.94	.93	3.46	1.30	484	2.94	1.10	2.31	1.18	517	4.37	.88	2.44	1.33	551	3.26	1.06	2.13	1.16
451	2.28	1.05	2.72	1.19	485	2.80	1.11	2.19	1.17	518	3.41	.93	2.46	1.13	552	1.59	.98	1.85	.92
452	3.57	1.13	2.33	1.30	486	1.89	.99	2.37	1.28	519	3.80	1.28	2.02	1.29	553	4.11	.91	2.21	1.21
453	2.94	1.20	2.08	1.99	487	2.22	1.11	2.06	1.21	520	2.30	1.05	2.85	.96	554	1.98	1.13	1.88	1.06
454	3.09	1.23	1.88	1.00	488	1.61	.86	2.12	1.13	521	2.24	1.02	2.41	1.09	555	3.91	.94	2.31	1.08
455	3.35	1.06	1.92	1.06	489	2.09	.99	2.58	1.19	522	1.96	1.07	2.35	1.25	556	2.15	1.01	2.48	1.24
456	3.61	1.00	2.23	1.37	490	1.78	1.01	2.23	1.20	523	2.07	.95	2.23	1.02	557	1.76	.99	1.81	1.01
457	3.09	1.07	2.04	1.20	491	3.07	1.14	1.90	.93	524	1.94	1.16	3.06	1.46	558	3.33	1.06	2.65	1.15
458	2.50	1.19	2.23	1.21	492	2.61	1.27	2.83	1.37	525	2.87	1.19	2.44	1.14	559	3.15	1.10	2.27	1.16
459	3.89	1.02	2.31	1.26	493	2.26	1.08	2.38	1.19	526	4.17	1.08	2.12	1.29	560	3.28	1.15	2.29	1.47
460	1.89	.99	2.10	1.27	494	3.91	.99	2.19	1.33	527	1.91	1.17	2.25	1.14	561	1.44	.89	1.81	.99
461	2.96	1.10	3.26	1.12	495	2.44	1.07	2.42	1.02	528	1.96	1.19	2.37	1.09	562	2.07	1.16	1.96	1.03
462	3.37	.97	2.20	1.31	496	1.87	.96	2.29	1.19	529	2.24	1.11	2.37	1.10	563	1.73	.94	1.85	.98
463	1.41	.72	1.83	.98	497	1.50	.75	2.13	1.05	530	3.35	1.04	2.08	1.12	564	1.80	.86	2.12	1.13
464	2.07	1.04	2.46	1.42	498	1.96	.99	2.42	1.11	531	3.35	1.06	2.31	1.18	565	4.30	1.03	2.29	1.47
465	2.22	1.09	2.21	1.23	499	3.02	1.18	2.44	1.16	532	2.26	1.22	2.19	1.01	566	2.13	1.19	2.46	1.13
466	1.65	.88	2.75	1.28						533	2.35	1.25	2.37	1.01					

Table 2
*Percentage of Items in Subtle-Obvious Categories for
 Minnesota Multiphasic Personality Inventory Clinical Scales*

Scale ^a	Very subtle	Somewhat subtle	Neither subtle nor obvious	Somewhat obvious	Very obvious
<i>Hs</i>	03	09	58	30	00
<i>D</i>	08	28	33	25	05
<i>Hy</i>	08	30	32	30	00
<i>Pd</i>	04	20	38	26	12
<i>Mf</i>	42	28	26	05	00
<i>Pa</i>	02	20	20	27	30
<i>Pt</i>	00	04	31	58	06
<i>Sc</i>	01	00	32	46	20
<i>Ma</i>	06	30	39	15	08
<i>Si</i>	12	34	35	17	00

^a The rating weight for an item was assigned within any particular scale in correspondence to the way that item was scored in the scale being considered. Therefore, if an item was scored false for that scale, then the false value was used, and if scored true for that scale, then the true value was used.

Table 1. It is interesting to note that when all MMPI items were rated as true, 333 of them were defined as subtle (i.e., mean value of less than 3.00) and 233 were defined as obvious (i.e., mean values greater or equal to 3.00).

Although mean values were calculated separately for males and females, space limitations allowed only for presentation of the combined ratings. It should be noted, however, that the correlations between male and female raters were very high; the product-moment correlations for true and false items were .94 and .90, respectively.

The combined mean obviousness ratings for the standard MMPI scales, ordered from most obvious to most subtle, were: *F* = 3.70; *Sc* = 3.64; *Pa* = 3.52; *Pt* = 3.47; *Hs* = 3.13; *Pd* = 3.13; *D* = 2.94; *Ma* = 2.82; *Hy* = 2.81; *Si* = 2.64; *L* = 2.41; *K* = 2.28; and *Mf* = 2.21. (The rating weight for an item was assigned within any particular scale in correspondence to the way an item was scored in the scale being considered. Therefore, if an item was scored false for that scale, then the false value was used, and if scored true, then the true value was used.)

With the exception of *Pa* and *D*, there is a perfect linear relationship between the mean obviousness scores for the clinical scales and the amounts of *K* correction added. This relationship between the suppressor *K* variable

and obviousness of item content of the scales is understandable. Although the ordering of the scales suggests that both *Pa* and *D*, but especially *Pa*, share the item characteristics associated with the need for the addition of *K*, McKinley, Hathaway, and Meehl (1948) found that the addition of *K* did not increase the predictive power of *D* and *Pa*. Explanations of this apparent anomaly are that (a) correction items were included on the *D* scale to minimize elevations on psychiatric cases whose primary diagnosis was not depression, and (b) more than 20% of the *Pa* items were subtle in character. (Cf. Table 2 for a comparison of *Pa*, *Pt*, and *Sc* with regard to percentages of very subtle and somewhat subtle items.) The correction items and the subtle items both function as *K*; hence, the addition of some fraction of *K* to these two scales was unnecessary, as *K* did not improve discriminative efficiency.

Table 2 shows the percentages of items in each of the subtle-obvious categories for each of the MMPI clinical scales. Examination of this table shows that *Pa* contains the highest percentage of items rated as "very obvious." However, if the two "obvious" categories are combined, one finds that *Sc* with 66% and *Pt* with 64% rank first and second on this dimension. Duff (1965) found that *Sc* contained a higher proportion of obvious items than *Hy* or *Pd*; percentages were 40, 22, and 6,

respectively. *Mf* contains far more items rated as "very subtle" (i.e., 42%) than any other scale. If the two "subtle" categories are combined, it can be seen that *Mf* ranks first and *Si* second in subtlety. There was also considerable variability in the neutral category; it is somewhat surprising to find that over half of the *Hs* items were considered neither obvious nor subtle, especially in view of the fact that this scale has been considered as a marker variable for obviousness (cf. Wiener, 1948).

The obviousness ratings for all subjects were compared with the desirability ratings of Messick and Jackson (1961). Since Messick and Jackson's ratings are for items answered true only, only our ratings of true responses were used. The product-moment correlation was $-.78$ ($p < .001$), which suggests that the subtle-obvious dimension and desirability have much in common. However, the relationship is not high enough to substitute desirability for obviousness ratings or vice versa.

Analysis of previous subtlety-obviousness measures in terms of the subtle-obvious ratings obtained in the present study demonstrated that the different procedures produced results that are superficially similar. Mean ratings of X and O items² were 3.16 and 2.04, respectively. Statistical analysis showed that these differences were highly significant, $t(565) = 3.58$, $p < .001$; however, it should be noted that the mean value of the X items was in our "neither subtle nor obvious" category. Wiener's (1948) subtle items ($M = 2.44$) were significantly more subtle than his obvious items ($M = 3.45$), $t(221) = 2.06$, $p < .05$, but neither set was extreme in terms of our ratings. The mean ratings of Duff's (1965) subtle, intermediate, and obvious items were 2.69, 3.38, and 3.60, respectively. Analysis of variance indicated that these means were significantly different, $F(2, 185) = 26.06$, $p < .001$. Further post hoc tests showed that the mean ratings of Duff's subtle items were significantly more subtle than the intermediate items, but the intermediate items were not significantly different from the obvious items. Here again it might be pointed out that the mean value of Duff's subtle items falls in our "neither subtle nor obvious"

category. The similarity then is more apparent than real, inasmuch as these previous rating schemes for the most part yielded values that tend to cluster in the neutral or obvious categories. Very subtle and very obvious items, in particular, seem to be strikingly underrepresented.

Jackson (1971) has argued that "those items are the best which are clearest and which contain referents about which people agree" (p. 239). Goldberg and Slovic (1967) compared responses to abstract designs with responses to items designed to measure achievement and affiliation and demonstrated that "only scales built from items of the highest face validity had significant cross-validity" (pp. 466-467). Duff (1965) found an inverse relationship between MMPI item subtlety and item discriminating power. Koss and Butcher (1973) showed that the MMPI items that characterized crisis situations displayed content clearly relevant to the particular situation (e.g., "The future seems hopeless to me" for the depressed-suicidal crisis). We have used these subtle-obvious ratings to show that subtle items are highly resistant to faking (Burkhart, Christian, & Gynther, 1978) and that endorsement of subtle items is characteristic of psychologically minded subjects (Burkhart, Gynther, & Christian, 1978). However, the most pressing problem is to determine the amount of trustworthy information provided by the endorsement of obvious and subtle items or keys. It may be that obvious/subtle keys for certain scales provide useful information, whereas obvious/subtle keys for other scales do not. In that case, one would want to use the valid obvious and subtle keys to maximize the discriminative power of the inventory. However, if further research confirms Duff's (1965) contention that subtle keys contribute no or very little information, development of a new inventory containing all obvious items (or an MMPI revised toward the same end) would be in order.

² These classifications are given in Appendix A of Dahlstrom, Welsh, and Dahlstrom (1972). It should be noted that the O items found in this appendix are not the same as the zero items used by Seeman in several studies (e.g., Wales & Seeman, 1969).

Reference Note

1. Wiener, D. N., & Harmon, L. R. *Subtle and obvious keys for the MMPI: Their development*. (Advisement Bulletin No. 16). Minneapolis, Minn.: Regional Veterans Administration Office, 1946.

References

- Burkhart, B. R., Christian, W. L., & Gynther, M. D. Item subtlety and faking on the MMPI: A paradoxical relationship. *Journal of Personality Assessment*, 1978, 42, 76-80.
- Burkhart, B. R., Gynther, M. D., & Christian, W. L. Psychological mindedness, intelligence, and item subtlety endorsement patterns on the MMPI. *Journal of Clinical Psychology*, 1978, 34, 76-79.
- Cronbach, L. J. *Essentials of psychological testing* (3rd ed.). New York: Harper & Row, 1970.
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook* (Vol. 1): *Clinical interpretation*. Minneapolis: University of Minnesota Press, 1972.
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook* (Vol. 2): *Research findings*. Minneapolis: University of Minnesota Press, 1975.
- Duff, F. L. Item subtlety in personality inventory scales. *Journal of Consulting Psychology*, 1965, 29, 565-570.
- Edwards, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden Press, 1957.
- Goldberg, L. R., & Slovic, P. Importance of test item content: An analysis of a corollary of the deviation hypothesis. *Journal of Counseling Psychology*, 1967, 14, 462-472.
- Harris, J. G., & Baxter, J. C. Ambiguity in the MMPI. *Journal of Consulting Psychology*, 1965, 29, 112-118.
- Heineman, C. E. Favorability ratings. Cited in W. G. Dahlstrom & G. S. Welsh, *An MMPI handbook*. Minneapolis: University of Minnesota Press, 1960. Pp. 431-433.
- Jackson, D. N. The dynamics of structured personality tests: 1971. *Psychological Review*, 1971, 78, 229-248.
- Koss, M. P., & Butcher, J. N. A comparison of psychiatric patients' self-report with other sources of clinical information. *Journal of Research in Personality*, 1973, 7, 225-236.
- McKinley, J. C., Hathaway, S. R., & Meehl, P. E. The MMPI: VI: The K scale. *Journal of Consulting Psychology*, 1948, 12, 20-31.
- Meehl, P. E. The dynamics of "structured" personality tests. *Journal of Clinical Psychology*, 1945, 1, 296-303.
- Meehl, P. E., & Hathaway, S. R. The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology*, 1946, 30, 525-564.
- Messick, S., & Jackson, D. N. Desirability scale values and dispersions for MMPI items. *Psychological Reports*, 1961, 8, 409-414.
- Wales, B., & Seeman, W. What do MMPI zero items really measure? An experimental investigation. *Journal of Clinical Psychology*, 1969, 25, 420-424.
- Wiener, D. N. Subtle and obvious keys for the Minnesota Multiphasic Personality Inventory. *Journal of Consulting Psychology*, 1948, 12, 164-170.
- Wiggins, J. S. An MMPI item characteristic deck. *Educational and Psychological Measurement*, 1964, 24, 137-141.
- Wiggins, J. S. Substantive dimensions of self-report in the MMPI item pool. *Psychological Monographs*, 1966, 80(22, Whole No. 630).

Received March 10, 1977 ■

Factor Structure of the SCL-90 in a Psychiatric Population

Norman G. Hoffmann and Peggy B. Overall
University of Texas Medical Branch, Galveston

The responses of an unselected psychiatric outpatient sample to the SCL-90, a self-report checklist of symptom complaints, were factor analyzed using a principal components procedure. Varimax rotation yielded five interpretable factors for which factor scores were derived. Results were compared with earlier studies using symptom checklists on selected outpatient diagnostic groups. Implications for future work with self-report symptom checklists are discussed.

Most clinical research in psychiatry over the past decade has relied on symptom and behavior ratings made by professional observers to describe psychopathology and to measure therapeutic response (Lorr, Klett, & McNair, 1963; Overall, 1974). There are numerous instances, however, for which it would seem desirable to assess how patients perceive their own states. The Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1967), the Psychological Screening Inventory (PSI; Lanyon, 1970), and other pencil-and-paper inventories have proved useful for description and classification of psychopathology, but as a general rule they have not appeared as useful as might be desired for the assessment of change. This is perhaps because such instruments are designed to measure more stable personality traits rather than states. A self-report psychiatric symptom checklist that was developed out of the more general item pool of the Cornell Medical Index (CMI) has been proposed as an instrument capable of measuring clinical states subject to therapeutic intervention and has undergone preliminary tests in psychopharmacology research. Originally called the Hopkins Symptom Check List (HSCL), the instrument has undergone sev-

eral revisions (Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1974).

Investigations have considered the factor structure of the original 58-item HSCL (Derogatis, Lipman, Covi, & Rickels, 1971, 1972; Lipman, Covi, Rickels, Uhlenhuth, & Lazar, 1968; Lipman Rickels, Covi, Derogatis, & Uhlenhuth, 1969; Mattsson, Williams, Rickels, Lipman, & Uhlenhuth, 1969; Williams et al., 1968). Four to six factors have been reported to account for the bulk of the variance. Although the labels assigned to the factors have varied, depression and anxiety factors have consistently emerged. Other factors frequently noted are represented by items dealing with somatic concerns, obsessive-compulsive themes, and interpersonal sensitivity.

The SCL-90, a 90-item revision of the HSCL, was created by the addition of 32 items concerned with symptoms of more serious psychopathology. Content of the added items concerns psychotic symptoms, paranoid ideation, phobic anxiety, and hostility (Derogatis, Lipman, & Covi, 1973). Thus, the SCL-90 would appear to have potential for use in the assessment of psychopathology in a more general psychiatric population.

Lipman, Covi, and Shapiro (1977) accomplished a factor analysis of the SCL-90 using data collected in a study involving chemotherapy of depressed patients. They identified eight factors, which they labeled as Interpersonal Sensitivity, Phobic Anxiety, Retarded Depression, Anger-Hostility, Somatization, Obsessive-Compulsive, Agitated Depression, and Psychoticism. Thus, they ob-

This study was supported in part by Department of Health, Education, and Welfare Grant 5 RO1 MH14675-09.

Requests for reprints should be sent to Norman G. Hoffmann, who is now at the Department of Psychiatry, St. Paul-Ramsey Hospital and Medical Center, St. Paul, Minnesota 55101.

tained factors similar to those found most often in analyses of the 58-item HSCL plus a psychoticism dimension defined predominantly by the added items.

A major concern that motivated the present investigation arises from the fact that the bulk of all previous factor analyses of the HSCL as well as the SCL-90 have focused on patients selected for symptoms of anxiety and/or depression. Diagnostic differences have been shown to affect the interpretation of factors on the HSCL (Derogatis et al., 1972), and similar results could be expected with the SCL-90.

This may be a particular problem when the intent is to use the SCL-90 for assessment of psychopathology in more general psychiatric populations. Factors that appear to be relatively independent in a restricted sample may be too highly correlated to be identified as separate factors in a more heterogeneous sample. Conversely, factors that are clearly present in a general psychiatric population may be absent in a highly selected, homogeneous sample. The purpose of the present article is to document the structure and statistical properties of the SCL-90 in a clinic population unselected for diagnosis and representative of the general psychiatric outpatient population seen in public facilities such as medical school clinics.

Method

The SCL-90 was administered to 358 new patients seen in the Adult Psychiatry Outpatient Clinic at

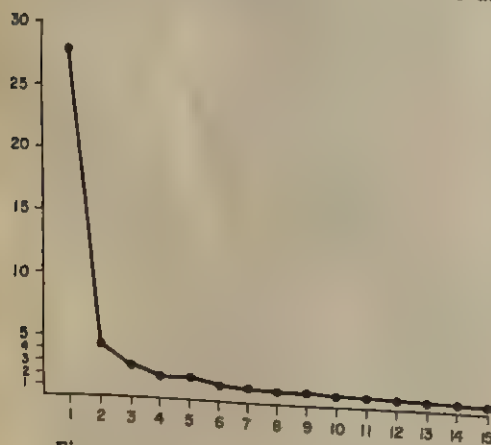


Figure 1. Eigenvalues for primary factors.

the University of Texas Medical Branch. The sample was unselected with the exception that an eighth-grade education and professed ability to read the newspaper was required. The examiner also heard the subjects read aloud the two sample items and respond to them. Verbal informed consent was obtained from each patient for the statistical research use of the test data. Demographic and diagnostic information was recorded to provide additional description of the sample.

The factor analysis used a principal components procedure, and the significant factors were rotated by the orthogonal normalized varimax method.

Factor scoring keys were defined to include the items that were found to project most highly on each of the rotated factors. To be included in a factor score, an item was required to have its highest loading on that factor. No items with a loading less than .350 were considered for factor scoring. Factor scores were then calculated for each subject by summing the values of the items included in the factor. The reliabilities of a single total score of pathology and the individual factor scores were estimated using coefficient alpha (Cronbach, 1951), which is an internal consistency statistic that assesses the average of all possible split-half coefficients. The intercorrelations among the factor scores were calculated to evaluate the extent to which the factors really represent distinct or independent aspects of psychopathology.

Inspection of the demographic variables indicates that the patients seen in this clinic are predominantly younger women from the lower socioeconomic classes. Almost 75% of the sample consisted of women, and over half of the patients were under 30 years of age. Ethnically, the group contained blacks and whites with a distinct minority of Mexican Americans who comprised 10% of the sample. An approximation of social class using educational and work levels in the "Two-Factor Index of Social Position" (Hollingshead, Note 1) resulted in almost 60% of the patients being classified into the lowest social class and only 13% of the sample placed into the three upper social classes.

In terms of psychopathology the current sample seems to differ substantially from those in previous studies of the HSCL and SCL-90. Although over 40% of the patients had a diagnosis of depression, many of those would not be candidates for inclusion in drug studies. Many cases involved personal crises as precipitating events. Indeed, 34% of all the patients seen were not prescribed any medication. Furthermore, the sample includes individuals with a variety of neurotic complaints, and personality problems as well as thought disorders, as is reflected in the fact that 16% of the patients were diagnosed as schizophrenic or schizoaffective.

Results

The pattern of eigenvalues obtained from the principal components analysis is shown in Figure 1. These results are presented to em-

phasize the prominence of a single general factor in the SCL-90 responses of patients in this outpatient clinic population. The first unrotated factor accounted for 6.45 times as much variance as the next largest and more than twice the variance of the next six factors combined. This suggests that in this patient population, the SCL-90 tends to measure a unitary global complaint factor and that the self-report complaints of the patients are not highly differentiated with respect to the more specific factors the instrument may be capable of measuring.

The display of the eigenvalues also suggests that five factors account for variance greater than the random scree level. The fifth factor is the first to depart in any measure from the nearly flat linear function that might be expected from random noise.

In point of fact, rotated solutions for 3-8 factors were examined, and the five-factor solution provided the most interpretable results. When only 3 or 4 factors were rotated, the content of the items loading on each factor appeared heterogeneous, making interpretation of the factors less clear. Conversely, when more factors are rotated, the later factors tend to fragment into interpretatively similar factors.

The five normalized varimax factors that were clearly defined and interpretable are Depression, Somatization, Phobic Anxiety, Functional Impairment, and Hostile Suspiciousness. Of the 90 items, 81 met the criteria of a loading greater than .350 and could be placed in one of the factors.

Examination of the item content found within each factor provides a straightforward picture of the basis for the interpretation of the factor. The factors identified here tend to agree well with those reported by Lipman et al. (1977). The bulk of the items in their two depression factors project highly on the single depression factor found here. The Somatization and Phobic Anxiety factors are almost identical to Lipman et al.'s in terms of the items with highest loadings on the respective factor. These three factors were found to be the most prominent in this unselected sample of patients.

The other factors, while not showing such a striking similarity to those obtained by previ-

ous investigators, do tend to have considerable overlap of items. The Functional Impairment factor in the present study most closely corresponds to the obsessive-compulsive factor in the Lipman et al. (1977) analysis. The bulk of the items in the Anger-Hostility and Psychoticism factors of the earlier study project on the Hostile Suspiciousness factor in the present study. The Interpersonal Sensitivity factor in the Lipman et al. study had highest loadings for six of its seven items evenly divided between the Depression and Hostile Suspiciousness factors described here, and it is the only Lipman et al. factor not related primarily to only one factor of the present study.

Factor scores were calculated by summing the responses to the sets of items indicated under each factor. Reliability of the factor scores was estimated using coefficient alpha, which is derived from internal consistencies among the item responses and is equivalent to the mean of all possible split-half reliability coefficients. The reliabilities for the five factor scores were .94, .90, .89, .83, and .94, respectively, indicating the high degree of consistency among the items that compose each factor.

The correlations between the factor scores were quite high, ranging from .504 to .747. This state of affairs is not unexpected in view of the very large eigenvalue associated with the first unrotated factor from the primary analysis and the fact that many items with their highest loading on one factor also had substantial projections on other factors. Even though it might have been possible to use only those items that projected onto only one factor, this would have greatly reduced the number of items available for the factor scales.

The total score on the SCL-90 was reliable and highly correlated with each of the factors, further suggesting that a single global score might well be used as an index of psychopathology or psychological discomfort. The correlations of the total score with the five factors were .928, .746, .831, .789, and .906, respectively. The Spearman-Brown split-half reliability between odd and even items was .976, and the alpha coefficient for the total test was .975.

For comparison, the factors defined by Lipman et al. (1977) were scored for this general outpatient sample. Intercorrelations among the eight Lipman factor scores ranged from .407 to .853, and the correlations of these factors with the total score ranged from .769 to .910. Again, it is apparent that at least some of the factors are not clearly distinguishable in the responses of patients in this general psychiatric clinic population. If one were to correct for attenuation due to measurement error, some of these correlations would approach unity. These results support our own in suggesting that scores that are ostensibly associated with distinct aspects of psychopathology are, in fact, measuring the same thing in this type of psychiatric clinic population.

Discussion

The rotated factors derived from the SCL-90 responses of a sample of unselected psychiatric outpatients appeared quite similar to the responses of depressed and/or anxious patients selected in previous psychopharmacology studies. These similarities were noted despite the fact that the current sample included a substantial proportion of individuals who were not deemed in need of medication, as well as more disturbed individuals who were taking antipsychotic medication. The checklist items and format would appear to tap some common dimensions in a broad range of patients.

The Depression, Somatization, and Phobic Anxiety factors seem to be the most clearly defined in the present study and are most consistent with other work on the HSCL and SCL-90. Of the three, the Somatization factor appears to be a slightly more independent dimension, in that its correlations with the other factors scores and the total test score tend to be lower.

Despite consistencies with factor structures in previous studies, the large proportion of variance accounted for by the first unrotated factor and the high intercorrelations among the factors raise some concerns that the instrument measures more of a general complaint or general discomfort dimension than distinct dimensions of psychopathology. Because of this, we have reservation about the

use of factor score profiles as has been proposed. More investigation is needed to determine whether factor scores of the SCL-90 do indeed produce profiles that provide a basis for differentiating different aspects of pathology.

Whether or not the SCL-90 can be used for specialized interpretations, such as distinguishing between anxiety and depression, our results indicate definite promise for the instrument as a global index of psychopathology or psychological distress. The total score derived from the summation of all items can be recommended for such a purpose. The inclusion of the SCL-90 in psychopharmacological studies seems quite appropriate for furthering the understanding of the properties of the instrument and exploring its construct validity. The evolution of better criterion measures is certainly needed in drug evaluation research, and the SCL-90 may hold promise as such a measure.

Reference Note

1. Hollingshead, A. B. *Two-factor index of social position*. Unpublished manuscript, Yale University, 1957.

References

- Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
- Derogatis, L. R., Lipman, R. S., & Covi, L. SCL-90: An outpatient psychiatric rating scale—Preliminary report. *Psychopharmacology Bulletin*, 1973, 9, 13-28.
- Derogatis, L. R., Lipman, R. S., Covi, L., & Rickels, K. Neurotic symptom dimensions: As perceived by psychiatrists and patients of various social classes. *Archives of General Psychiatry*, 1971, 24, 454-464.
- Derogatis, L. R., Lipman, R. S., Covi, L., & Rickels, K. Factorial invariance of symptom dimensions in anxious and depressive neuroses. *Archives of General Psychiatry*, 1972, 27, 659-665.
- Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., & Covi, L. The Hopkins Symptom Checklist (HSCL): A measure of primary symptom dimensions. In P. Pichot (Ed.), *Psychological measurement: Modern problems in pharmacopsychiatry* (Vol. 7). Basel, Switzerland: Karger, 1974.
- Hathaway, S. R., & McKinley, J. C. *The MMPI manual* (Rev. ed.). New York: The Psychological Corporation, 1967.
- Lanyon, R. I. Development and validation of a Psychological Screening Inventory. *Journal of Consulting and Clinical Psychology*, 1970, 35, 1-24.

- Lipman, R. S., Covi, L., Rickels, K., Uhlenhuth, E. H., & Lazar, B. A. Selected measures of change in outpatient drug evaluation. In, *Psychopharmacology: A review of progress 1957-1967* (Public Health Service Publication, No. 1836). Washington, D.C.: U.S. Government Printing Office, 1968.
- Lipman, R. S., Covi, L., & Shapiro, A. K. The Hopkins Symptom Check List (HSCL): Factors derived from the HSCL-90. *Psychopharmacology Bulletin*, 1977, 13, 43-45.
- Lipman, R. S., Rickels, K., Covi, L., Derogatis, L. R., & Uhlenhuth, E. H. Factors of symptoms distress: Doctor ratings of anxious neurotic patients. *Archives of General Psychiatry*, 1969, 21, 328-338.
- Lorr, M., Klett, C. J., & McNair, D. M. *Syndromes of psychosis*. New York: Pergamon Press, 1963.
- Mattsson, N. B., Williams, H. V., Rickels, K., Lipman, R. S., & Uhlenhuth, E. H. Dimensions of symptom distress in anxious neurotic outpatients. *Psychopharmacology Bulletin*, 1969, 5, 19-32.
- Overall, J. E. The Brief Psychiatric Rating Scale in psychopharmacology research. *Modern Problems of Pharmacopsychiatry*, 1974, 7, 67-78.
- Williams, H. V., et al. Replication of symptom distress factors in anxious neurotic outpatients. *Multivariate Behavior Research*, 1968, 3, 199-212.

Received April 6, 1977 ■

Notice on Author Alterations

Effective January 1979, the procedure for billing authors for alterations will change. Authors of articles in APA journals are customarily charged for alterations on proofs that result from (a) authors' errors or omissions in the manuscript and (b) changes that result from authors' failure to review the edited manuscript or the proof. Until now, alteration charges have been determined by the number of lines affected by changes requested and have been based on 1974 printing rates. Beginning with the proofs for the first issues of 1979, authors' alterations will be determined not only by the number of lines affected but also by the number of figures and pages affected and will be based on current printers' rates.

This change in alteration billing is being made so that alteration charges will more closely reflect the actual cost of changes requested and may mean larger alteration bills for some authors. Authors are reminded that alteration charges apply only when changes are made on proofs and that *there is no charge* for changes made at the manuscript stage.

Internal-External Expectancies and Health-Related Behaviors

Bonnie R. Strickland
University of Massachusetts, Amherst

The present article is a review of research on internal-external (I-E) locus of control expectancies and health attitudes and behaviors. The theoretical background of the I-E construct is described. Topics covered include I-E in relation to health knowledge, precautionary health practices, reactions to physical disorders, psychological responding, psychological disturbances, and responses to psychological treatment. Some problems and issues are also noted.

Health concerns and health costs constitute a social problem of enormous magnitude in this country (APA Task Force on Health Research, 1976). Estimates of numbers of individuals with symptoms severe enough to warrant attention with respect to treatment range as high as 75% of the general population (Mechanic, 1972). The cost of medical care is the fastest growing item in the U.S. consumer price index, close to \$100 billion annually, and it is expected to more than double over the next 5 years ("Health-Cost Crisis," 1977). For good reason, individuals appear to be increasingly concerned not only with recognizing and attending to debilitating symptoms but also preventing the occurrence of illness or accident. The rewards to those persons who remain free of disease or disability are, of course, not only financial but include enhanced physical and emotional well-being as well. The findings from a broad range of studies demonstrating a generally greater adaptive functioning for those persons holding internal as opposed to external expectancies (Strickland, Note 1, Note 2) have clear implications with respect to health.

Earlier versions of this article have been presented at meetings of the American Psychological Association in 1972 and 1973 and at the XV InterAmerican Congress in 1974. The writing of these articles and the present one were supported in part by funds available through the National Institute of Child and Human Development, Grant 08952.

Requests for reprints should be sent to Bonnie R. Strickland, Department of Psychology, University of Massachusetts, Amherst, Massachusetts 01003.

The internal-external control of reinforcement (I-E) dimension is an expectancy variable couched within Rotter's social learning theory (Lefcourt, 1976; Phares, 1976; Rotter, 1954; Rotter, Chance, & Phares, 1972; Strickland, 1977). Simply stated, I-E refers to the degree to which an individual perceives the events that happen to him/her as dependent on his/her own behavior or as a result of luck, chance, fate, or powers beyond one's personal control and understanding. Assessment of I-E expectancies is via questionnaires, of which the Rotter I-E scale (Rotter, 1966) has been the instrument of choice for most ongoing research with adults. Several multidimensional instruments to assess I-E have been devised (Collins, 1974; Gurin, Gurin, Lao, & Beattie, 1969; Levenson & Miller, 1976), and some researchers have developed I-E measures specific to health (Kirscht, 1974; Wallston, Wallston, Kaplan, & Maides, 1976). Results of research conducted with the various instruments suggest that beliefs about internal versus external control are related in significant and even dramatic ways to health-related behaviors. The purpose of the present article is to review these relationships across several broad areas of health, ranging from prevention to susceptibility to remediation of physical and psychological dysfunctions.

The Theoretical Framework

The implication that I-E expectancies are related to the facilitation of health behaviors

is derived from social learning theory. Rotter (1954; Rotter et al., 1972) postulated behavior to occur as a function of expectancy and reinforcement within a specific situation. If a situation is novel or ambiguous, then an individual will depend on generalized expectancies that have served him/her in the past. More specific expectancies are used when the aspects of the situation are straightforward or routine. The I-E dimension is a *generalized* expectancy that occurs when individuals have learned that events are contingent or non-contingent on their behavior. Individuals holding internal expectancies are more likely than externals to take responsibility for their actions (Davis & Davis, 1972; Phares, Wilson, & Klyver, 1971) and to attribute responsibility to agents who activate chance (Hochreich, 1972; Phares & Wilson, 1972; Schiavo, 1973; Sosis, 1974). In performance task situations, internals are perceptually alert and attentive (DuCette & Wolk, 1973; Lefcourt, Gronnerud, & McDonald, 1973; Lefcourt & Wine, 1969; Wolk & DuCette, 1974) and appear to gather and process information effectively for problem solving (Davis & Phares, 1967; DuCette & Wolk, 1972; Pines & Julian, 1972). Research on social action (Gore & Rotter, 1963; Levenson & Miller, 1976; Pawlicki & Almquist, 1973; Sanger & Alger, 1972; Strickland, 1965) suggests that individuals who believe that events are related to their own behaviors are more likely than persons trusting fate or powers beyond their control to take steps to change aversive life situations. Phares (1976, p. 78) proposed that the cognitive and motivational aspects of the I-E dimension lead internals to a superior position in exerting power and control over their environment. If this is the case, then I-E expectancies may have significant impact in relation to health maintenance, a most important personal concern for many of us. Our parents and ourselves are enjoying a longer life span, with the accompanying infirmities of age, than has been true in the past. Even in good health, we are daily bombarded by the public media with information about the likelihood of disease occurring from such ordinary events as breathing city air and eating bacon. We are constantly urged to improve our health by losing weight, jogging, and engaging in all

those inviting, energetic activities designed to enhance our physical functioning. When we do experience physical or emotional distress, professional health care is more readily available to us than it has been in the past, and most of our friends and acquaintances have advice and their own favorite home remedies to share. One would expect that internals, in contrast to externals, would be more sensitive to health messages, would have increased knowledge about health conditions, would attempt to improve physical functioning, and might even, through their own efforts, be less susceptible to physical and psychological dysfunction.

Health Knowledge and Precautionary Measures

As early as 1962, Seeman and Evans found evidence that hospitalized patients with tuberculosis who were internal as assessed by an early I-E measure, with intelligence controlled, knew more about their disease than their matched external counterparts. The medical staff also rated internal patients higher in objective knowledge about tuberculosis than externals. In wards in which information was difficult to obtain, internal patients were significantly less satisfied with the flow of information than externals. The Wallstons and their colleagues (Wallston, Maides, & Wallston, 1976; Wallston, Wallston, Kaplan, & Maides, 1976) have also found that internals who value their health are more likely than others to collect information about disease and health maintenance when alerted to possible hazards, such as hypertension.

Following the growing concern in the 1960s about a link between cigarette smoking and cancer, numerous studies were conducted which suggested that individuals who were not smokers or individuals who were able to stop smoking were more internal than individuals who smoked (Coan, 1973; James, Woodruff, & Werner, 1965; Mlott & Mlott, 1975; Steffy, Meichenbaum, & Best, 1970; Straits & Sechrest, 1963; Williams, 1973; Platt, Note 3). These results have not always been replicated (Danaher, 1977; Lichtenstein & Keutzer, 1967), but taken together they do suggest that individuals with internal rather than ex-

ternal expectancies are more likely to take action to improve their health habits, particularly when faced with evidence that needed changes may result in improved physical functioning.

In a study of inoculation against influenza, Dabbs and Kirscht (1971) reported that college students who were assessed as internal on eight "motivational" variables were more likely than externals to have been inoculated, although internals on eight "expectancy" variables were not likely to have taken the shots. In other words, subjects who said that they were motivated to exert control in relation to health behavior did indeed take precaution against susceptibility to an infectious disease. Subjects who simply reported expectancies about contingencies between behavior and events were not differentiated according to having received flu shots. In a large-scale study with high school students, Williams (1972a) found that internal students reported greater use of seat belts when riding in an automobile than externals. These students also reported themselves to be significantly more likely to engage in preventive dental care, that is, to go to the dentist for checkups and maintenance even when teeth or gums are not sore or hurting (Williams, 1972b). Sonstroem and Walker (1973) found internal college males to hold more positive attitudes toward physical exercise and cardiovascular fitness than externals; these internal students were also more likely to participate in voluntary exercise.

In other research designed to investigate preventive health practices, Balch and Ross (1975) tested 34 females and reported internal beliefs to be predictive of both success in and completion of an overweight treatment program. Wallston, Wallston, Kaplan, and Maides (1976) found 22 female subjects in weight reduction programs consistent with their specific I-E beliefs about health to be more satisfied with treatment. Results of weight loss for subjects in congruent conditions were in the expected direction, although they failed to reach significance. Other experimenters have not been able to relate I-E to attempts at weight loss (Bellack, Rozen-sky, & Schwartz, 1974; Manno & Marston, 1972; Tobias & MacDonald, 1977), although

some reported overweight subjects to be external, a finding also reported by O'Bryan (1972).

At least two studies suggest that females who are internal are more likely than external females to practice birth control effectively (Lundy, 1972; MacDonald, 1970). However, Harvey (1976) assessed almost 200 female college undergraduates as to their reliance on safe (pill, intrauterine devices) versus risky (condom, rhythm, etc.) birth control methods and found no differences as a function of I-E expectancies. Segal and DuCette (1973) reported that middle-class white high school females who became pregnant were more external than control subjects; lower class, black, high school females who became pregnant were more internal. The authors suggested that pregnancy has different meanings for these populations. At that time, in the upper-class group, pregnancy might have been an undesirable event, and internal students would have been expected to take precautions against pregnancy. Pregnancy may have been more acceptable among the lower-class group and for some students might even have been a source of pride and success. If these assumptions are correct, then the discrepant I-E predictions make sense. Obviously, a number of complex factors enter into any decision, or lack thereof, about pregnancy. However, the I-E variable appears to be a promising one in relation to predicting the use of birth control, especially if goals of family planning and rationale for use of birth control can be specified.

With some exceptions, the bulk of the reported research on I-E and precautionary health practices lends credence to the expected theoretical assumptions that individuals who hold internal as opposed to external expectancies are more likely to assume responsibility for their health. Internals appear to attempt to maintain their physical well-being and to guard against accidents and disease to a greater extent than individuals who hold external expectancies. As would be expected within Rotter's theory, internals *who value their health* (reinforcement value) seek more information about health maintenance, and when stricken with a disorder

they appear to learn more about the disease that afflicts them. Other specific and generalized expectancies as well as situational contingencies would also be expected to interact with what is likely a complex relationship between I-E and precautionary health practices.

Reactions to Physical Disorders

When a person is faced with a chronic handicap or a debilitating illness, do I-E expectancies play any part in an individual's reaction to this situation? MacDonald and Hall (1971) asked healthy college students how they might respond to various physical handicaps in regard to social relationships and feelings about themselves. External rated physical disabilities as more debilitating than did the internals, who apparently anticipated less severe consequences of handicaps. But, what happens in real life? Lipp, Kolstoe, James, and Randall (1968) investigated the responses of handicapped and normal subjects to pictures of disabled people. The handicapped subjects suffered a variety of disabilities including amputation, paralysis, fractures, arthritis, and congenital deformities. Normal subjects were of approximately the same age and sex as the handicapped group and were matched on the basis of responses to the James I-E scale. Slides of disabled and normal persons were presented tachistoscopically to all subjects. As the investigators had hypothesized, disabled subjects took significantly more trials to recognize the disability slides than the normal subjects. Of interest to the present article are their findings of an interaction between three levels of I-E and the disabled/nondisabled dichotomy. The external disabled individuals were less denying of disability, as measured by recognition time of disability slides, than the internals, including those subjects whose scores fell in the middle I-E ranges. The investigators suggested that the internals are more threatened and hence more denying of disability than the externals.

Several other studies have been conducted investigating I-E in relation to threat to ascertain if internals and externals do respond differently. Phares, Ritchie, and Davis

(1968) provided subjects with either positive or negative feedback, which ostensibly resulted from their answers to a number of personality tests. Consistent with a denial model, externals, as assessed by the Rotter scale, recalled significantly more of the evaluative material, both positive and negative, than did internals. Internals did, however, demonstrate a greater willingness to engage in remedial behavior to confront their alleged problems. Another investigation by Houston (1972) gives additional information about the differential influence of I-E beliefs in regard to response to stress. Houston told one group of subjects that they could avoid electric shock in an experimental condition by not making mistakes on a subtest of the Wechsler Adult Intelligence Scale. Another group was told that there was no way to avoid shock. Subjects who perceived that they had some control over the shock reported less anxiety but actually evidenced greater physiological arousal than the no-control group. Although they did not report more distress, internal subjects showed an increase in heart rate significantly greater than did externals across both conditions. Evidently the internals were more aroused in the control of shock condition, but they denied their anxiety about the situation. Externals may have found it easier to "accept" the threat of shock and resign themselves to the situation. These results, of course, raise implications about the relation of I-E beliefs to a host of psychological findings on "perceived control" (Glass & Singer, 1972). I-E beliefs may have direct relevance, for example, to the Type A and Type B behavior patterns that Glass (1977) described. Glass wrote that Type A individuals, who sound strikingly like internals, appear to be engaged in a struggle for control and try to develop strategies for coping with uncontrollable stress. This continued attempt at mastery of life events, however, takes a tragic toll in that these individuals appear to be more prone to coronary heart disease.

Aside from studies on reactions of individuals to perceived threat, numerous investigations with persons who have actually suffered some traumatic event have been conducted. Blind children (Land & Vineberg,

1965), children with cerebral palsy (Egglund, 1973), and children with severe reading problems (Strickland & Hill, Note 4) have all been assessed as more external than comparable, nonhandicapped children. Jones (1974), however, investigated orthopedically disabled children and found no relationship between I-E beliefs and degree of mobility. Wendland (1973) tested 80 males, ages 18-35, with muscular skeletal impairment. Subjects who had been disabled less than 1½ years were significantly more external than subjects disabled 3 years or longer. Wendland suggested that disabled individuals have a tendency to expect increased direction from external forces during the initial period following disability onset.

Bruhn, Hampton, and Chandler (1971) compared a group of 36 male hemophilics, ages 12 and over, with a control group of normals. They found that overall, the hemophilic group was more internal than normal controls. But, within the hemophilic group, a marginally severe group was significantly more external than either a mild or a severe group. These investigators suggested that the marginally severe hemophiliac views his clinical state as unpredictable and is more dependent on external cues to determine his well-being. Goldstein (1976) also compared 24 long-term male hemodialysis patients with 22 male patients, all of whom were recovering from minor medical problems. The hemodialysis sample patients obtained significantly higher denial and externality scores than the nonhemodialysis control subjects.

A number of studies that are investigations of attempts to influence health care, once disabilities occur, are available. Weaver (1972) found that internal patients with severe kidney disorders who were using dialysis machines to stay alive were significantly more likely than matched externals to comply with diet restrictions and to keep scheduled appointments. Internal patients hospitalized with spinal cord injury had higher self-concepts and reported themselves to be less depressed than matched externals (Dinardo, 1972). Those internals who scored high on the repression side of Byrne's Repression-Sensitization scale showed the best adjustment, and external sensitizers showed the

poorest adjustment to spinal cord injury. However, Bulman and Wortman (1977) found no differential I-E predictions in a sample of 24 paraplegics or quadraplegics. Ireland (1973) attempted to investigate participation in treatment in relation to I-E beliefs among pulmonary emphysema patients. Ratings were difficult to obtain, and no clear findings emerged. He did find, however, as noted earlier, that internal patients knew more about their disorder than externals, even with intelligence controlled.

Presurgical external patients report more anxiety than matched internals (Lowery, Jacobsen, & Keane, 1975), and, following abdominal surgery, internal female patients are more likely to attempt to influence postoperative care in relation to obtaining more analgesics than external patients (Johnson, Leventhal, & Dabbs, 1971). If they were firstborns, internals also had longer hospital stays than externals. Johnson et al. explained the interaction between I-E and birth order as follows: They postulated that firstborns who are exposed to stress become frightened and dependent and so they try to remain in contact with the authorities who control danger. Internal firstborns have learned to manipulate the health personnel in order to remain in the hospital. External firstborns do not have sufficient skill to influence postoperative stay. Since later born patients are less dependent and frightened, they feel no need to stay longer regardless of I-E beliefs.

In another rather complicated research design, Auerbach, Kendall, Cuttler, and Levitt (1976) found no relationship between I-E scores and anxiety about impending dental procedures among a group of patients scheduled for surgical removal of a tooth. However, they did find that internal and external patients responded differentially to specific and general presurgery information. According to dentist ratings, internals adjusted poorly in surgery after receiving general, marginally relevant material about the dental procedure. However, internals showed good adjustment in surgery after viewing a tape that imparted specific information about the procedures and sensations that they might expect; the reverse was true for external patients. These investigators suggested

that internals responded favorably to the specific information, since it provided relevant input consistent with their cognitive set that they exert control of the occurrence of reinforcers and punishments provided them. The authors hypothesized that the specific information provided data that enhanced the internal's perception that he/she might manipulate the impending aversive situation, whereas the general information reinforced the ambiguity of the situation and a lack of personal control over it. They suggested that for externals, the specific information led to a diminished reliance on outside sources as the precipitating events leading to this aversive situation. The general tape allowed the external patient to avoid personal responsibility, an action congruent with a defensive posture about control of reinforcement.

In one of the most complex and well-controlled studies of personality characteristics in relation to stress, heart attack, and recovery, Cromwell, Butterfield, Brayfield, and Curry (1977) manipulated nursing care, participation in various activities, and information about heart attack for 229 coronary patients. Eighty medical patients with illnesses comparable in severity but without cardiovascular involvement served as controls. Dependent variables included stay in intensive care, stay in hospital, rate of alarms (heart rate changes while on unit), a number of biochemical and physiological indices, rehospitalization, and death. Overall, in regard to I-E scores as assessed by the Rotter scale, coronary patients were more external than the medical controls. No patients who were involved in congruent combinations of locus of control beliefs and participation in self-treatment (internals with high participation and externals with low participation) returned to the hospital ($p < .06$) or died ($p < .06$) within 12 weeks following their hospital stay. The small number of patients who did return to the hospital ($n = 12$) or who died ($n = 5$) had all been involved in incongruent conditions. As indicated by the .06 significance levels, these findings cannot be accepted with great confidence. However, the findings are in the predicted direction, and the small number of cases could have been contributed to the lack of statistical

significance. Since these are long-term effects, Cromwell et al. suggested that patients whose hospital treatment was incongruent with personal expectancies may have resisted a decision to return to the hospital when crucial symptoms of another myocardial infarction appeared. I-E beliefs were also related to a number of other dependent variables. Except in two cases in which internality was linked with high anxiety, externality was always associated with undesirable physical characteristics such as higher temperature and higher sedimentation rates. Also, internal patients were more cooperative in response to treatment demands, and they left the coronary unit and the hospital earlier than external patients. It should be noted, however, that Marston (1969) found no relationship between I-E beliefs and compliance behavior among a group of coronary patients and, with 58 male patients, Garrity (1973) found that subjects' perceptions of health status, social class, and a belief in external control predicted return to work after their first myocardial infarction.

Aside from the Cromwell et al. (1977) study, very little research is available linking I-E beliefs to specific physical illnesses. However, Naditch (1974) considered data on over 400 black men and women (in six American cities) who were diagnosed as having essential hypertension—a "silent" health hazard for large numbers of black people. Among externals who rated themselves as discontented with their lives, the rate of hypertension was 46%, more than double the 21% rate for the total sample and considerably higher than rates for all other groupings (e.g., a 7% rate for contented internals). In further analysis, Naditch concluded that these results occurred primarily as a function of the responses of males but not of females in the sample. Similar to Cromwell et al., and as would be expected from the Glass (1977) research, the Naditch I-E findings have implications for cardiovascular involvement. Taken together, these studies suggest that I-E beliefs may be particularly salient for understanding the influence of belief about control on physiologically adaptive responses to stress within the cardiovascular system. (See Strickland, in press,

for a more complete review of I-E and cardiovascular functioning.)

Darrow (1973) tested several hundred persons who presented themselves to community health centers for diagnosis and treatment of venereal disease. He found men assessed as internal to be significantly less likely to be infected with gonorrhea. No such results emerged from females, although internal women were more likely to return for follow-up treatment with the appearance of new symptoms than were external females. Darrow interpreted this latter finding as most likely being due to the tendencies of these women to notice physiological changes and to seek an explanation for the reoccurrence of symptoms after treatment. Olbrisch (1975) also reported a complex relationship between externality and naive beliefs about gonorrhea, in that external subjects appear to have a more casual, helpless attitude about how venereal diseases are contracted. In this study, however, externals did not differ from internals in plans to make further precautions.

Any impending or disabling disorder, whether chronic or temporary, has a varying degree of influence on the responses of the persons faced with the handicap. The severity of the disorder, the time of onset, the current status of the patient, the support that he/she receives, and so on, all interact with what is probably a complex set of cognitions about the disorder. When an individual is more helpless than he/she once was, or is handicapped in relation to others, beliefs about locus of control would be expected to be, and apparently are, related to reactions to the disorder and the struggle to recover. Although chronically handicapped individuals, particularly children, appear to be more external than their control counterparts, internal adults, in spite of the fact that they may respond to disablement with initial denial and concern, appear to know more about their disorder and attempt to influence health care to a greater extent than externals. These results are most clear in those complex designs that give attention to the interactions of I-E expectancies and situational demands. For example, internals seems to be able to use specific information about their disease

and treatment, whereas externals respond to general instructions. Given the initial defensiveness and denial of internals faced with trauma, a number of questions are raised as to when internal versus external expectancies are more adaptive. It may be that a defensive stance is helpful when a person who is accustomed to considerable personal control is suddenly faced with events beyond his or her influence.

I-E and Physiological Responses

Aside from the findings of relationships of I-E expectancies and general health practices, it is of interest to note the degree to which I-E beliefs may be related to an individual's ability to monitor and change specific physiological responses. A number of investigators and health care personnel have attempted to teach individuals to improve physical functioning via the use of biofeedback, and indications of individual differences in response to this technique become particularly important. Again, a logical assumption is that persons who hold strong locus of control expectancies, whether internal or external, will have differing responses to attempts to control their own internal physical states. Internals would be expected to be sensitive to internal states, alert to biofeedback cues, and motivated to attempt self-control of bodily function.

Results of several studies do show internals to be generally superior to externals in responding to biofeedback paradigms. Internals are better able to increase and maintain electroencephalogram alpha responding (Gosling, May, Lavond, Barnes, & Carreira, 1974; Johnson & Meyer, 1974) and to lower galvanic skin responses via biofeedback than externals (Wagner, Bourgeois, Levenson, & Denton, 1974). Other biofeedback research also demonstrates the influence of I-E expectancies on the control of vascular responses. Ray (1974) found internals to be more proficient at increasing heart rate and externals at decreasing heart rate than comparison subjects. These results were essentially replicated by Gatchel (1975) for the first training trial of a biofeedback paradigm. Fotopoulous (1971) reported internal sub-

jects to be more capable of increasing heart rate without either reinforcement or external feedback, whereas externals could increase heart rate only under a reinforcement paradigm.

Like the clinical findings on I-E/cardiovascular relationships, these data on the voluntary control of heart rate suggest that internals and externals may be using different strategies in biofeedback paradigms and that effective responding might be enhanced if individuals are in conditions that are congruent with their expectancies for control. As noted by Cromwell et al. (1977), internals may respond to opportunities to work individually, and externals may need conditions of structure or outside influence to enhance their responding. At least one laboratory study suggested that this is indeed the case. DeGood (1975) studied 24 internal and 24 external subjects under one of two aversive shock avoidance procedures. Half of the subjects could escape shock by asking for a rest period, and half had rest periods imposed by the experimenter. Control over initiation of rest had an arousal-reducing effect on systolic blood pressure for all subjects. Diastolic blood pressure change appeared to be a function of an interaction of I-E expectancies and the situation. Elevations were lowest when personal and situational control factors were congruent, that is, for internals in conditions of self-initiation of rest and for externals under imposed rest.

In a particularly complex biofeedback design, Carlson (in press) studied 24 male and 24 female college students who equally represented three distinct ethnic groups: Caucasians, Japanese, and Chinese. The feedback subjects acquired lower frontal electromyographic (EMG) levels than control subjects, and internal subjects in the feedback condition acquired lower levels than externals. No consistent differences in EMG levels were obtained in the control condition as a function of I-E (as assessed by the Nowicki-Strickland I-E scale for adults; Nowicki & Strickland, 1973). These results were stable across both sexes, all three ethnic groups, and across two replications. On pretest and posttest, Carlson found that external subjects in the feedback conditions shifted significantly

toward internality, suggesting that the experiencing of the opportunity for changing bodily states was related to an enhanced internal expectancy. No change scores occurred for the internal subjects, possibly because of a ceiling effect. Internal subjects in the control conditions did shift toward externality, but this change was not statistically significant. Additionally, I-E shifts were apparently not related to actual performance changes. Despite the fact that internals in the feedback condition achieved the lowest frontal EMG levels, they also reported feeling somewhat less relaxed during training than their counterparts in the control condition and the externals in the feedback conditions. Carlson suggested that in their efforts to perform well in a frontal muscle relaxation task, internals may actually sacrifice their subjective state of general relaxation. Again, these findings are reminiscent of the cardiovascular results in which internals appear to deny arousal while actually experiencing distress in situations in which they have no control.

Berggren, Ohman, and Fredrikson (1977) shed some additional light on the possible mechanisms that lead to the differential physiological responses of internals and externals to stimulus conditions. When college students with extreme I-E scores were asked to respond to a recurring tone of moderate intensity, external subjects took significantly longer to reach a criterion of habituation than internals on a measure of skin conductance. Evidently, externals were exhibiting continued electrodermal orienting responses to nonsignal stimuli. The investigators then ran additional subjects in the same methodological procedure plus a second experimental manipulation—The stimulus was given more significance by arranging that it would be a signal for a forthcoming task. Some groups of subjects were asked to press a switch at the *offset* of a recurring tone, and other groups repeated Experiment 1. Again, externals took significantly longer to reach criterion of habituation than did internals. In the signal condition, however, this effect was reversed so that internals habituated more slowly. The external group failed to differentiate between signal and nonsignal conditions. The investigators interpreted these results to suggest that

externals have poorer control of attention than internals. External subjects attended to irrelevant events and did not seem to differentiate between relevant and irrelevant cues. Internals, on the other hand, differentiated sharply between cues and stopped responding to irrelevant cues quickly. Overall, as theoretically expected, internals appeared to be vigilant, with more active attentional processes and more focus on task-relevant cues.

Physiological responding in biofeedback designs appears to be functionally related to I-E expectancies and enhanced in congruent conditions. It is likely that these effects occur because internals and externals attend and respond to relevant stimuli in different ways. Internals appear to be more motivated to succeed and are generally superior in performance than externals. However, their vigilance and attempts to control their bodily states may result in immediate increased arousal and diminished future well-being. Again, it is important to know when internals should be encouraged to relax and/or to relinquish control for enhanced physical functioning. These investigations could have direct and practical implications for preventive health practices, especially in regard to cardiovascular disease. Both internals and externals can learn to control heart rate. However, they use different strategies to do so, and their efforts may lead to different long-term outcomes. More specific investigations are needed, but evidently cognitive mediating variables about perceived control are impactful in relation to basic cardiovascular functioning.

I-E and Psychological Disturbances

Research on the I-E variable and the reporting of psychological and/or emotional difficulties is much more extensive than that on I-E and physical disorders. At a general level of overall functioning, internal individuals including the elderly (Felton & Kahana, 1974; Wolk & Kurtz, 1975) are significantly more likely to report themselves as content with their life situations than externals (Naditch, Gargan, & Michael, 1975; Palmore & Luikart, 1972). The relationships among I-E and adjustive behavior and attitudes, however, is moderated by the nature of the set-

tings in which people reside (Wolk, 1976). With regard to dysfunctional difficulties, investigators have found a belief in external locus of control to be related to debilitating anxiety (Butterfield, 1964; Feather, 1967; Finch & Nelson, 1974; Platt & Eisenman, 1968; Shriberg, 1974; Strassberg, 1973; Watson, 1967), to the holding of irrational values (MacDonald & Games, 1972), to mood disturbances (Kilpatrick, Dubin, & Marcotte, 1974), and to indices of maladjustment on paper-and-pencil questionnaires (Duke & Nowicki, 1973; Hersch & Scheibe, 1967; Powell & Vega, 1972). With patients who have been hospitalized for psychiatric reasons, a number of researchers have reported a relationship between externality and severity of psychiatric diagnosis (Cash & Stack, 1973; Croft, Johnson, & Fox, 1975; Cromwell, Rosenthal, Shakow, & Zahn, 1968; Duke & Mullins, 1973; Harrow & Ferrante, 1969; Lefcourt, 1976; Levenson, 1973; Lottman & DeWolfe, 1972; Palmer, 1971; Shybut, 1968; C. E. Smith, Pryer, & Distefano, 1971). These data are correlative, and there is no way of knowing if external beliefs accompany a predisposition to psychological difficulties or if locus of control beliefs occur as a function of the disturbances. At the least, it appears that the reporting of life contentment is related to internality, whereas pathological difficulties are linked to external expectancies.

A puzzling issue throughout a consideration of I-E and maladaptive behavior, however, concerns the discrepant predictions about externality and depression. One might logically expect that individuals who believe that they are responsible for the results of their behavior would be more likely to become depressed when life events do not go well for them than persons who are able to attribute traumatic events to luck, fate, God's judgment, and so forth. Indeed, Phares (1972) has hypothesized that "depressions tend to be associated with people who possess a strong generalized expectancy that outcomes are their own responsibility" (p. 466). The guilt and self-punitiveness often expressed by depressives would be expected to occur only if individuals actually believe that they influence life occurrences. On the other hand, many depressives report themselves to be powerless

about life events, to experience loss of control, and to feel helpless about influencing future events—all perceptions that sound similar to externality. Strickland (Note 1) and Lefcourt (1976) have both hypothesized a relationship between depression and externality, and much of the empirical literature supports this contention (S. I. Abramowitz, 1969; Calhoun, Cheney, & Dawes, 1974; Dinardo, 1972; Emmelkamp & Cohen-Kettenis, 1975; Goss & Morosko, 1970; Moyal, 1977; Naditch et al., 1975; Prociuk, Breen, & Lusier, 1976; Wareheim & Foulds, 1971; Haley & Strickland, Note 5). Considerable evidence has also accumulated with regard to I-E beliefs in relation to "learned helplessness"—a phenomenon proposed by Seligman (Abramson, Seligman, & Teasdale, 1978; Seligman, 1974, 1975) as a model for reactive depression. Generally, researchers find that externals show poorer performance in the learned helplessness paradigm than do internals (Cohen, Rothbart, & Phillips, 1976; Hiroto, 1974). Although, in at least one instance, individuals responded to helplessness manipulations with increased attempts at control (Roth & Bootzin, 1974). In fact, Brehm's reactance theory (Brehm, 1966, 1972) would predict that a person threatened with loss of freedom will become motivationally aroused to prevent their loss. Obviously, individuals have differing coping styles and respond to circumstances in diverse but possibly predictable ways. Increased research with the I-E variable might give additional clues as to individual responses to aversive or traumatic life situations. Research in this area must remain completely open, however, without presupposed judgments of what is "good" or "adaptive." As Wortman and Brehm (1975) caution, an emphasis on personal causation may be dangerous when individuals are faced with situations that are truly uncontrollable. Additional complications in this kind of research have to do with the difficulty of defining depressive responses. Depression is a multidimensional disorder, and facets of depression (e.g., reactive vs. chronic depression) may be differentially related to I-E. In fact, Strickland and Hale (Note 6) found external expectancies to be more strongly related to a measure of chronic depression than to temporary depressed mood.

Further, when factors or dimensions within the I-E construct are examined, they lead to differential predictions. For example, depressives take responsibility for negative but not positive events in their lives (Haley & Strickland, Note 5). Clearly, the relationships are complicated and further research is necessary. The I-E variable would appear to be particularly useful in relation to cognitive theories of depression (Beck, 1976) and in models incorporating response-contingent hypotheses (Lewinsohn, 1972). Of particular importance are those critical studies linking cognitive variables to overt behavior.

Even though depression takes its toll in lessened feelings of well-being and happiness as well as in loss of energy, perhaps the most serious aspect of depression is the possibility of suicide that occurs for many depressed individuals. Several investigators have considered the relationships among suicidal thought, behavior, and I-E expectancies in an attempt to better understand life-threatening behaviors. Lambley and Silbowitz (1973) could not predict the contemplation of suicide via Rotter's scale, but Williams and Nickels (1969), in a study of 235 college students, found externality to be related to suicide potential as measured by the Minnesota Multiphasic Personality Inventory. Crepeau (Note 7) also asked college students how often they contemplated suicide. Generally, he found suicide ideation to be linearly related to Collins' (1974) I-E measure, with persons who reported suicidal thoughts being assessed as more external than students who had never considered suicide.

Melges and Weisz (1971) talked with 15 patients who had recently made serious suicide attempts. They asked them to recall as vividly as possible the feelings and thoughts that they were experiencing immediately before the attempt. With pretest and posttest measures, they found increased externality following the specific suicide ideation. Patients also reported more negative evaluations of the future and less extension of a span of awareness toward the future. Both of these variables were also related to changes toward externality. Thus, findings from reports of the reexperiencing of suicidal thoughts suggest that the original suicidal impulses may

reflect a feeling of loss of control and an inability to foresee a pleasant future.

Although many of the empirical results reported in psychological literature point to a relationship between beliefs about external control and psychopathology, this finding does not hold in some selected samples of maladaptively functioning persons. For example, considerable confusion is apparent in studies with individuals who are substance abusers. Some investigators reported alcoholics to be more external than control samples (Butts & Chotlos, 1973; Naditch, 1975; Nowicki & Hopper, 1974; Obitz & Swanson, 1976; Palmer, 1971), and others found alcoholics to be more internal (Goss & Morosco, 1970; Oziel, Obitz, & Keyson, 1972). Caster and Parsons (1977) found that alcoholics score higher than control subjects on the Chance Control dimension of Levenson's scale and noted that treatment success varied in relation to I-E scores. A correlation between external Control by Powerful Others and depression was evident in successfully treated groups, whereas a relationship between depression and Chance Control was found for failures. Caster and Parsons suggested that alcoholics who are depressed and who see their distress as occurring because of powerful others may respond better than those alcoholics for whom depression is psychologically related to fate or chance. Other investigators have suggested that I-E differentiates among alcoholics, with those holding internal expectancies being less severely disturbed. Internal alcoholics also experience a greater magnitude of control over intrapersonal and interpersonal stresses than the external alcoholics (Donavan & O'Leary, 1975; Donovan, O'Leary, & Schau, 1975; O'Leary, Donovan, & Hague, 1974a, 1974b; Pryer & Distefano, 1977). Although one might theoretically expect drug abusers to be external, some investigators reported them to be more internal than control groups (Berzins & Ross, 1973; Calicchia, 1974; Smithyman, Plant, & Southern, 1974). Again, a number of factors may be operating here, with complications arising because of the complexities of the disorders, difficulties in diagnosis, and the impact of situational demands in the testing sessions. Rotter (1975) noted that alcohol and drug abusers may be responding to the

exhortations of staff members in treatment programs who tell patients that the success of treatment is "up to them." They may dissimulate on I-E measures to present themselves in a favorable light. Although Rotter (1966) attempted to control for social desirability in the early I-E assessment instruments, a number of investigators have reported a relationship between internality and socially desirable responses (Cone, 1971; Harris, 1975; Hjelle, 1971; Vuchinich & Bass, 1974). Thus, some of the reported relationships between externality and the reporting of psychopathological symptoms may be a function of approval-motivated response set rather than veridical representation of locus of control expectancies.

I-E and Psychological Treatment

Assuming for the moment, however, that externality and maladaptive functioning are related raises a number of questions about responses to psychotherapy or treatment programs in relation to I-E expectancies. Conceptually, internals might enter therapy or treatment at a less disturbed level than externals. Once in treatment, internals would be expected to respond in adaptive ways, assuming responsibility for their difficulties and attempting to change. This assumption, of course, depends on the therapeutic approach (Nowicki, Bonner, & Feather, 1972). Internals might be quite resistant to interventions that they perceive as limiting their freedom or control. And, as mentioned earlier, therapeutic benefits might be most enhanced when individuals are in treatment situations congruent with their locus of control beliefs.

Generally, results of a large number of studies across different treatment modalities suggest that individuals in therapy or self-improvement groups do become more internal as treatment progresses (Diamond & Shapiro, 1973; Dua, 1970; Eitzen, 1974; Gillis & Jessor, 1970; Lewis, Dawes, & Cheney, 1974; Kilmann & Howell, 1974; Lynch, Ogg, & Christensen, 1975; Pierce, Schauble, & Farakas, 1970; Schallow, 1975; R. E. Smith, 1970). Academic underachievers also appear to become more internal in response to counseling and structured group activities (Fel-

ton & Biggs, 1972, 1973; Felton & Davidson, 1973; Felton & Thomas, 1972; Nowicki & Barnes, 1973; Reimanis, 1974), as do juvenile felons (Moser, 1975).

Several more studies demonstrate the more complicated interactive effects when subjects are differentiated according to I-E expectancies and placed in varied treatments (Abramowitz, Abramowitz, Roback, & Jackson, 1974; Friedman & Dies, 1974; Kilmann, 1974; Kilmann, Albert, & Sotile, 1975; Morley & Watkins, 1974). Generally, if given an opportunity, internals report that they prefer more client control than do externals. Internals respond more positively to nondirective approaches in which therapist intervention is minimal and structure is not imposed from the outside. Externals, on the other hand, appear more positively influenced by structured approaches. These findings parallel those behavior modification results that occur when individuals are differentiated as to I-E expectancies and exposed to different procedures (Best, 1975; Wallston, Wallston, Kaplan, & Maides, 1976). For example, Best and Steffy (1975) involved internals and externals in one of two smoking modification procedures. Congruence of I-E expectancies and experimental conditions produced the most profound changes. Internals responded to an aversion satiation procedure, and externals responded to an agent who decided the rate at which smoking would be reduced.

Aside from responses of clients or patients, several investigators have assessed I-E expectancies of mental health personnel and have suggested that locus of control beliefs may be important in determining the efficacy of mental health care delivery. Beckman (1972) found volunteers in a state mental hospital to be more internal than undergraduate controls and to be less likely to believe that patients should be restricted in their social functioning. Felton (1973) reported mental health workers to be more internal as a function of a process-oriented training program. Martin and Shepel (1974) trained 21 senior female nurses from urban hospitals in counseling, emphasizing developing a helping relationship, identifying and exploring problem areas, and devising plans of action. Through pretesting and posttesting, they

found both an increase in counseling perceptiveness and a shift toward internality. A correlation of .56 was found between counseling skills and internality at the posttesting.

Overall, the reported research for both individuals in treatment programs and individuals delivering mental health care suggests that internal expectancies may facilitate responsible adaptive responses. Results appear to be highly influenced, however, by the conditions under which counseling or treatment occurs. Again, congruence between locus of control expectancies and the structure of the therapeutic endeavor appears to lead to the most pervasive changes.

Problems

The increase in research on I-E expectancies and health-related behaviors is quite striking since the first Seeman and Evans study in 1962. The blending of interest in individual difference variables and health most likely results from an increased focus on preventive health care in this country and a changing social and political awareness in which individual responsibility for one's own health is emphasized. Certainly, a number of health practitioners have noted that many of the psychological and physical disorders that persons bring into a physician's or counselor's office result from, or are exacerbated by, behavior such as smoking, improper diet, lack of exercise, and substance abuse. Others have remarked on the striking individual differences in response to treatment programs once problems or physical dysfunctions are identified. Finally, increasing health costs have caused massive concern on the part of individuals who can no longer afford health care and the politicians who represent them. Overall, research with the I-E dimension suggests that beliefs about locus of control of reinforcement are influential in relation to health.

Some major problems with this research should be noted, however. First, much research that does not produce clear-cut or substantial results is not published. Implications of the published results are important, but these would, of course, be attenuated if they reflected only a small part of the work that has been done but not published.

A second major problem has to do with the relative magnitude of the I-E/health relationships that are reported. The I-E variable is only one of a number of complex factors that may converge to predict health attitudes and behaviors. The amount of variance for which I-E accounts is probably quite small in many, if not most, situations. For example, whether persons present themselves to a physician's office for relief of symptoms may be much more a function of severity of symptoms, the individual's financial condition, and/or the availability of health care than the person's beliefs about control of reinforcement. Thus, practical prediction and clinical intervention must be continually accompanied by further investigation of the complex of components that accompany health behaviors.

One of the most pervasive problems of research with the I-E dimension has to do with the differing definitions that have been attached to this construct and the fact that I-E expectancies are often, even usually, explored outside the theoretical net in which this concept was first described. Investigations using the I-E variable have proliferated, sometimes without a clear grounding in a precise understanding of the construct and its implications. Rotter (1966, 1975) has always noted that I-E is only one of a number of variables that would be expected to predict behavior in specific and novel situations. An equal concern with the nature of the situational demands and reinforcement value should improve prediction.

Aside from theoretical concerns, a number of methodological weaknesses are apparent in much of the I-E/health research. Particularly with the use of clinical populations, controls are often lacking with respect to severity and length of disorder or illness. Also, treatment methods vary when disorders are identified. Moreover, much of the research is correlational in nature and gives no indication of direction of causality. The findings that do emerge demand further investigation as to their continued viability and validity, with particular emphasis on antecedent conditions.

Another major methodological problem has to do with the measurement of I-E expectancies. Numerous I-E assessment instruments are currently being used. Most of the

research surveyed in the present article, unless otherwise noted, has been conducted with the Rotter scale, which has been the instrument in most frequent use. However, I-E measures different from the Rotter scale were sometimes used, particularly in the very earliest and the latest studies. The conflicting I-E/health findings that emerge might then be explained as occurring because of the use of the different assessment instruments that may not be measuring the same expectancies. Moreover, some I-E assessment devices are more appropriate for and better predictors of health concerns than others. The scale developed by the Wallstons, for instance, is specific to health and would be expected to enhance prediction of health-related behaviors.

Finally, the problem of response set is always one that haunts any self-report measure. This is especially telling for the I-E dimension, since significant correlations between the various I-E measures and social desirability responding are often reported. It is difficult to know the degree to which I-E responses reflect veridical descriptions of locus of control beliefs or are colored by the respondent's attempt to present himself/herself in a favorable light. Particularly in situations with strong social demand characteristics, such as treatment programs for incarcerated substance abusers in which responses may influence length of stay in institutions, individuals may be expected to respond in ways that are designed to please the controlling agents. Even within a more typical outpatient situation, persons may wish to present themselves favorably to the authorities and therapists responsible for their care, implying that they are motivated, concerned, conscientious, and so forth. Also, response bias may occur on the various I-E measures as a function of respondents' social-cultural background (Nowicki & Strickland, 1973), race (Gurin et al., 1969), sex (Strickland & Haley, Note 8), and political ideology (MacDonald, 1972). More accurate assessment of the degree and specificity of I-E expectancies, independent of mediator variables and response bias, if possible, would be enormously helpful.

In spite of the problems, research on the I-E dimension in relation to health appears to have opened significant avenues of investi-

gation that should be pursued. Although results are not altogether as clear, convincing, and as free of conflict as one might hope, the bulk of the research is consistent in implying that when faced with health problems, internal individuals do appear to engage in more generally adaptive responses than do externals. These range from engagement in preventive and precautionary health measures through appropriate remedial strategies when disease or disorder occurs. Findings suggest that the development of an internal orientation could lead to improved health practices for some individuals who have been inclined to believe that life events are beyond their responsibility and more a function of external control. One must be quite cautious, however, in assuming that internal beliefs are always facilitative. The continued alertness of internals and their attempts at mastery behavior is most appropriate when events are actually controllable. When individuals persist in efforts that bring no relief, then they may find themselves to be actually exacerbating the undesirable characteristics of the situation in which they find themselves. Perhaps the wisest course is that people learn to specify the reality of their life situations, their possible responses, and the potentiality of forthcoming reinforcement.

Another major finding emerging from this review is that congruence of expectancies and situations appears to enhance behavior change. Practical implications are that change agents such as health personnel will be most effective when techniques are tailored to individual expectancies. External individuals evidently respond more easily to conditions in which structure is imposed from outside. Internals prefer situations in which they can assume responsibility and work independently.

Obviously, continued research is necessary. The problems extant in the investigations under review are myriad, but the already observed results hold promise for both theoretical and practical advances.

Reference Notes

1. Strickland, B. R. *Locus of control: Where have we been and where are we going?* Paper presented at the meeting of the American Psychological Association, Honolulu, Hawaii, August 1973.

2. Strickland, B. R. *Locus of control and health-related behaviors*. Paper presented at the XV InterAmerican Congress of Psychology, Bogota, Columbia, December 1974.
3. Platt, E. S. *Internal-external control and changes in expected utility as predictions of the change in cigarette smoking following role playing*. Paper presented at the meeting of the Eastern Psychological Association, April 1969.
4. Strickland, B. R., & Hill, J. N. *An investigation of some personality variables in male children with severe reading problems*. Unpublished paper, University of Massachusetts, 1974.
5. Haley, W. E., & Strickland, B. R. *Locus of control and depression*. Paper presented at the meeting of the Eastern Psychological Association, Boston, April 1977.
6. Strickland, B. R., & Hale, W. D. *Dimensions of locus of control and self-reported depression*. Unpublished manuscript, University of Massachusetts, Amherst, 1975.
7. Crepeau, J. *The effects of stressful life events and locus of control on suicidal ideation*. Paper presented at the meeting of the Eastern Psychological Association, Boston, April 1977.
8. Strickland, B. R., & Haley, W. E. *Sex differences in the Rotter IE scale*. Unpublished manuscript, University of Massachusetts, Amherst, 1977.

References

- Abramowitz, C. V., Abramowitz, S. I., Roback, H. B., & Jackson, C. Differential effectiveness of directive and nondirective group therapies as a function of client internal-external control. *Journal of Consulting and Clinical Psychology*, 1974, 42, 849-853.
- Abramowitz, S. I. Locus of control and self-reported depression among college students. *Psychological Reports*, 1969, 25, 149-150.
- Abramson, L. Y., Seligman, M. E. P., & Teasdale, J. D. Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, 1978, 87, 49-74.
- American Psychological Association Task Force of Health Research. Contributions of psychology to health research. *American Psychologist*, 1976, 31, 263-274.
- Auerbach, S. M., Kendall, P. C., Cuttler, H. F., & Levitt, N. R. Anxiety, locus of control, type of preparatory information, and adjustment to dental surgery. *Journal of Consulting and Clinical Psychology*, 1976, 44, 809-818.
- Balch, P., & Ross, A. W. Predicting success in weight reduction as a function of locus of control: A unidimensional and multidimensional approach. *Journal of Consulting and Clinical Psychology*, 1975, 43, 119.
- Beck, A. T. *Cognitive therapy and the emotional disorders*. New York: International Universities Press, 1976.
- Beckman, L. Locus of control and attitudes toward mental illness among mental health volunteers.

- Journal of Consulting and Clinical Psychology*, 1972, 38, 84-89.
- Bellack, A. S., Rozensky, R., & Schwartz, J. A. Comparison of two forms of self-monitoring in a behavioral weight reduction program. *Behavior Therapy*, 1974, 5, 523-530.
- Berggren, T., Ohman, A., & Fredrickson, M. Locus of control and habituation of the electrodermal orienting response to nonsignal and signal stimuli. *Journal of Personality and Social Psychology*, 1977, 35, 708-716.
- Berzins, J. I., & Ross, W. F. Locus of control among opiate addicts. *Journal of Consulting and Clinical Psychology*, 1973, 40, 84-91.
- Best, J. A. Tailoring smoking withdrawal procedures to personality and motivational differences. *Journal of Consulting and Clinical Psychology*, 1975, 43, 1-8.
- Best, J. A., & Steffy, R. A. Smoking modification procedures for internal and external locus of control clients. *Canadian Journal of Behavioral Science*, 1975, 7, 155-165.
- Brehm, J. W. *A theory of psychological reactance*. New York: Academic Press, 1966.
- Brehm, J. W. *Responses to loss of freedom: A theory of psychological reactance*. Morristown, N.J.: General Learning Press, 1972.
- Bruhn, J. G., Hampton, J. W., & Chandler, B. C. Clinical marginality and psychological adjustment in hemophilia. *Journal of Psychosomatic Research*, 1971, 15, 207-213.
- Bulman, R. J., & Wortman, C. Attributions of blame and coping in the "real world": Severe accident victims react to their lot. *Journal of Personality and Social Psychology*, 1977, 35, 351-363.
- Butterfield, E. C. Locus of control, test anxiety, reactions to frustration, and achievement attitudes. *Journal of Personality*, 1964, 32, 298-311.
- Butts, S. V., & Chotlos, J. A. A comparison of alcoholics and nonalcoholics on perceived locus of control. *Quarterly Journal of Studies on Alcohol*, 1973, 34, 1327-1332.
- Calhoun, L. G., Cheney, T., & Dawes, A. S. Locus of control, self-reported depression, and perceived causes of depression. *Journal of Consulting and Clinical Psychology*, 1974, 42, 736.
- Calicchia, J. P. Narcotic addiction and perceived locus of control. *Journal of Clinical Psychology*, 1974, 30, 499-504.
- Carlson, J. G. Locus of control and frontal electrographic response training. *Biofeedback and Self-Regulation*, in press.
- Cash, T. F., & Stack, J. J. Locus of control among schizophrenics and other hospitalized psychiatric patients. *Genetic Psychology Monographs*, 1973, 87, 105-122.
- Caster, D. U., & Parsons, O. Relationship of depression, sociopathy, and locus of control to treatment outcomes in alcoholics. *Journal of Consulting and Clinical Psychology*, 1977, 45, 751-756.
- Coan, R. W. Personality variables associated with cigarette smoking. *Journal of Personality and Social Psychology*, 1973, 26, 86-104.
- Cohen, S., Rothbart, M., & Phillips, S. Locus of control and the generality of learned helplessness in humans. *Journal of Personality and Social Psychology*, 1976, 34, 1049-1056.
- Collins, B. E. Four components of the Rotter internal-external scale: Belief in a difficult world, a just world, a predictable world, and a politically responsive world. *Journal of Personality and Social Psychology*, 1974, 29, 381-391.
- Cone, J. D. Social desirability scale values and ease of responding to personality statements. *Proceedings of the 79th Annual Convention of the American Psychological Association*, 1971, 6, 119-120. (Summary)
- Croft, R. G., Johnson, W. G., & Fox, S. H. Locus of control in psychiatric patients. *Newsletter for Research in Mental Health and Behavioral Sciences*, 1975, 17, 17-20.
- Cromwell, R. L., Butterfield, E. C., Brayfield, F. M., & Curry, J. L. *Acute myocardial infarction: Reaction and recovery*. St. Louis, Mo.: Mosby, 1977.
- Cromwell, R. L., Rosenthal, D., Shakow, D., & Zahn, T. P. Reaction time, locus of control, choice behavior, and description of parental behavior in schizophrenic and normal subjects. *Journal of Personality*, 1968, 29, 363-380.
- Dabbs, J. M., & Kirscht, J. P. "Internal control" and the taking of influenza shots. *Psychological Reports*, 1971, 28, 959-962.
- Danaher, B. G. Rapid smoking and self-control in the modification of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1068-1075.
- Darrow, W. W. Innovative health behavior: A study of the use, acceptance, and use-effectiveness of the condom as a venereal disease prophylactic. (Doctoral dissertation, Emory University, 1973). *Dissertation Abstracts International*, 1973, 34, 2792A. (University Microfilms No. 73-25, 393)
- Davis, W. L., & Davis, D. E. Internal-external control and attribution of responsibility for success and failure. *Journal of Personality*, 1972, 35, 547-561.
- Davis, W. L., & Phares, E. J. Internal-external control as a determinant of information seeking in a social influence situation. *Journal of Personality*, 1967, 35, 547-561.
- DeGood, D. E. Cognitive control factors in vascular stress responses. *Psychophysiology*, 1975, 12, 399-401.
- Diamond, M. J., & Shapiro, J. L. Changes in locus of control as a function of encounter group experience. *Journal of Abnormal Psychology*, 1973, 82, 514-518.
- Dinardo, Q. E. Psychological adjustment to spinal cord injury. (Doctoral dissertation, University of Houston, 1971). *Dissertation Abstracts International*, 1972, 32, 4206B-4207B. (University Microfilms No. 71-27,248)
- Donavan, D. M., & O'Leary, M. R. Comparison of perceived and experienced control among alcoholics and nonalcoholics. *Journal of Applied Psychology*, 1975, 89, 726-728.
- Donavan, D. M., O'Leary, M. R., & Schau, E. J. Relationship between generalized expectancies for

- control and experienced control among alcoholics. *Psychological Reports*, 1975, 37, 71-76.
- Dua, P. S. Comparison of the effects of behaviorally oriented action and psychotherapy re-education on intraversion-extraversion, emotionality, and internal-external control. *Journal of Counseling Psychology*, 1970, 17, 567-572.
- DuCette, J., & Wolk, S. Locus of control and extreme behavior. *Journal of Consulting and Clinical Psychology*, 1972, 39, 253-258.
- DuCette, J., & Wolk, S. Cognitive and motivational correlations of general expectancy of control. *Journal of Personality and Social Psychology*, 1973, 26, 420-426.
- Duke, M. P., & Mullins, C. Interpersonal distance as a function of locus of control in hospitalized schizophrenics and nonschizophrenics. *Journal of Consulting and Clinical Psychology*, 1973, 41, 230-234.
- Duke, M. P., & Nowicki, S. Personality correlates of the Nowicki-Strickland Locus of Control Scale for Adults. *Psychological Reports*, 1973, 33, 267-270.
- Egglund, E. T. Locus of control and children with cerebral palsy. *Nursing Research*, 1973, 22, 329-333.
- Eitzen, D. S. Impact of behavior modification techniques on locus of control of delinquent boys. *Psychological Reports*, 1974, 35, 1317-1318.
- Emmelkamp, P. M., & Cohen-Kettenis, P. Relationship of locus of control to phobic anxiety and depression. *Psychological Reports*, 1975, 36, 390.
- Feather, N. T. Some personality correlates of external control. *Australian Journal of Psychology*, 1967, 19, 252-260.
- Felton, G. S. Teaching internalization to middle-level mental health workers in training. *Psychological Reports*, 1973, 32, 1279-1282.
- Felton, G. S., & Biggs, B. E. Teaching internalization to collegiate low achievers in group therapy. *Psychotherapy: Theory, Research, and Practice*, 1972, 9, 281-283.
- Felton, G. S., & Biggs, B. E. Psychotherapy and responsibility: Teaching internalization to black low achievers through group therapy. *Small Group Behavior*, 1973, 4, 147-155.
- Felton, G. S., & Davidson, H. R. Group counseling can work in the classroom. *Academic Therapy*, 1973, 8, 461-468.
- Felton, B., & Kahana, E. Adjustment and situationally-bound locus of control among the institutionally aged. *Journal of Gerontology*, 1974, 29, 295-301.
- Felton, G. S., & Thomas, L. J. How to beat the failure syndrome: A process-oriented learning program for collegiate low achievers. *College Student Journal Monograph*, 1972, 6, 1-13.
- Finch, A. J., & Nelson, W. Locus of control and anxiety in emotionally disturbed children. *Psychological Report*, 1974, 35, 469-470.
- Fotopoulos, S. Internal vs. external control: Increase of heart rate by thinking under feedback and no-feedback conditions. (Doctoral dissertation, University of Kansas, 1970) *Dissertation Abstracts International*, 1971, 31, 3703B-3704B. (University Microfilms No. 70-25,335)
- Friedman, M. L., & Dies, R. R. Reactions of internal and external test-anxious students to counseling and behavior therapies. *Journal of Consulting and Clinical Psychology*, 1974, 42, 921.
- Garrity, T. F. Vocational adjustment after first myocardial infarction: Comparative assessment of several variables suggested in literature. *Social Science and Medicine*, 1973, 7(9), 705-717.
- Gatchel, R. J. Change-over training sessions of relationships between locus of control and voluntary heart-rate control. *Perceptual and Motor Skills*, 1975, 40, 424-426.
- Gillis, J. F., & Jessor, R. Effects of brief psychotherapy on belief in internal control: An exploratory study. *Psychotherapy: Theory, Research, and Practice*, 1970, 7, 135-136.
- Glass, D. C. Stress, behavior patterns, and coronary disease. *American Scientist*, 1977, 65, 177-187.
- Glass, D. C., & Singer, J. E. *Urban stress: Experiments on noise and social stressors*. New York: Academic Press, 1972.
- Goldstein, A. M. Denial and external locus of control as mechanisms of adjustment in chronic medical illness. *Essence*, 1976, 1, 5-22.
- Gore, P. M., & Rotter, J. B. A personality correlate of social action. *Journal of Personality*, 1963, 31, 58-64.
- Gosling, W. J., May, C., Lavond, D., Barnes, T., & Carreira, C. Relationship between internal and external locus of control and the operant conditioning of alpha through biofeedback training. *Perceptual and Motor Skills*, 1974, 39, 1339-1343.
- Goss, A., & Morosko, T. E. Relation between a dimension of internal-external control and the MMPI with an alcoholic population. *Journal of Consulting and Clinical Psychology*, 1970, 34, 189-192.
- Gurin, P., Gurin, G., Lao, R., & Beattie, M. Internal-external control in the motivational dynamics of Negro youth. *Journal of Social Issues*, 1969, 25, 29-53.
- Harris, W. G. *An investigation of the EI scale's predictability using the bogus pipeline*. Unpublished master's thesis, University of Massachusetts, 1975.
- Harrow, M., & Ferrante, A. Locus of control in psychiatric patients. *Journal of Consulting and Clinical Psychology*, 1969, 33, 582-589.
- Harvey, A. A. Risky and safe contraceptors: Some personality factors. *Journal of Psychology*, 1976, 92(1), 109-112.
- Health-cost crisis. *Newsweek*, May 9, 1977, pp. 84-90.
- Hersche, P. D., & Scheibe, K. E. Reliability and validity of internal-external control as a personality dimension. *Journal of Consulting Psychology*, 1967, 31, 609-613.
- Hiroto, D. S. Learned helplessness and locus of control. *Journal of Experimental Psychology*, 1974, 102, 187-193.

- Helle, L. A. Social desirability as a variable in the locus of control scale. *Psychological Reports*, 1971, 24, 807-816.
- Hochreich, D. J. Internal and external control and reaction to the My Lai court martials. *Journal of Applied Social Psychology*, 1972, 2, 319-325.
- Houston, H. K. Control over stress, locus of control, and response to stress. *Journal of Personality and Social Psychology*, 1972, 22, 249-255.
- Ireland, R. E. Locus of control among hospitalized pulmonary emphysema patients (Doctoral dissertation, University of Kansas, 1972). *Dissertation Abstracts International*, 1973, 33, 6091A (University Microfilms No. 73-11,900).
- James, W. H., Woodruff, A. B., & Werner, W. Effect of internal and external control upon changes in smoking behavior. *Journal of Consulting Psychology*, 1965, 29, 184-186.
- Johnson, J. E., Leventhal, H., & Dabbs, J. M. Contribution of emotional and instrumental response processes in adaptation to surgery. *Journal of Personality and Social Psychology*, 1971, 20, 55-64.
- Johnson, R. K., & Meyer, R. G. The locus of control construct in EEG alpha rhythm feedback. *Journal of Consulting and Clinical Psychology*, 1974, 42, 913.
- Jones, R. L. Correlates of orthopedically disabled children's school achievement and interpersonal relationships. *Exceptional Children*, 1974, 41, 191-192.
- Kilmann, P. R. Direct and nondirect marathon group therapy and internal-external control. *Journal of Consulting Psychology*, 1974, 21, 380-384.
- Kilmann, P. R., Albert, R. M., & Sotile, W. M. Relationships between locus of control, structure of therapy, and outcome. *Journal of Consulting and Clinical Psychology*, 1975, 41, 588.
- Kilmann, P. R., & Maxwell, R. J. Effects of structure of marathon group therapy and locus of control on therapeutic outcome. *Journal of Consulting and Clinical Psychology*, 1974, 42, 917.
- Küster, D. G., Cohen, W. R., & Marcotte, I. B. Personality, stress, of the medical education process, and changes in affective mood state. *Psychological Reports*, 1974, 34, 1215-1223.
- Kursh, J. P. Perceptions of control and health beliefs. *Canadian Journal of Behavioral Science*, 1970, 4, 125-137.
- Landry, P., & Schwartz, M. Rotter's internal-external scale and prediction of suicide contemplations among students. *Psychological Reports*, 1973, 23, 149-150.
- Leach, S., & Vinberg, V. F. Locus of control in blind children. *Exceptional Children*, 1965, 31, 111-116.
- Levenson, H. W. Locus of control. *Current trends in theory and research*. Hillsdale, N.J.: Erlbaum, 1975.
- Levenson, H. W., Ginzburg, P., & McDonald, P. Cognitive biases and hypothesis formation during a simple stimulus word association test as a function of locus of control and field dependence. *Canadian Journal of Behavioral Science*, 1973, 5, 161-173.
- Lefcourt, H. M., & Wine, J. Internal versus external control of reinforcement and the deployment of attention in experimental situations. *Canadian Journal of Behavioral Science*, 1969, 1, 167-181.
- Levenson, H. Multidimensional locus of control in psychiatric patients. *Journal of Consulting and Clinical Psychology*, 1973, 41, 397-404.
- Levenson, H., & Miller, J. Multidimensional locus of control in sociopolitical activist of conservative and liberal ideologies. *Journal of Personality and Social Psychology*, 1976, 33, 199-208.
- Lewinsohn, P. M. Clinical and theoretical aspects of depression. In H. E. Adams, & W. K. Boardman (Eds.), *Advances in experimental psychology*. New York: Pergamon Press, 1972.
- Lewis, P., Dawes, A. S., & Cheney, T. Effects of sensitivity training on belief in internal control of interpersonal relationships. *Psychotherapy: Theory, Research and Practice*, 1974, 11, 282-284.
- Lichtenstein, E., & Keutzer, C. S. Further normative and correlation data on the internal-external (IE) control of reinforcement scale. *Psychological Reports*, 1967, 21, 1014-1016.
- Lipp, L., Kolstoe, R., James, W., & Randall, H. Denial of desirability and internal control of reinforcement: A study using a perceptual defense paradigm. *Journal of Consulting and Clinical Psychology*, 1968, 32, 72-75.
- Lottman, T. J., & DeWolfe, A. S. Internal versus external control in reactive and process schizophrenia. *Journal of Consulting and Clinical Psychology*, 1972, 39, 344.
- Lowery, B., Jacobsen, B., & Keane, A. Relationship of locus of control to reoperative anxiety. *Psychological Reports*, 1975, 37, 1115-1121.
- Lundy, J. R. Some personality correlates of contraceptive use among unmarried female college students. *Journal of Psychology*, 1972, 80, 9-14.
- Lynch, M. L., Ogg, W. D., & Christensen, M. G. Impact of a life planning workshop on perceived locus of control. *Psychological Reports*, 1975, 37, 1219-1222.
- MacDonald, A. P. Internal-external locus of control and the practice of birth control. *Psychological Reports*, 1970, 27, 206.
- MacDonald, A. P. More on the Protestant Ethic. *Journal of Consulting and Clinical Psychology*, 1972, 39, 116-122.
- MacDonald, A. P., & Games, R. G. Ellis' irrational values. *Rational Living*, 1972, 7, 25-28.
- MacDonald, A. P., & Hall, J. Internal-external locus of control and perception of desirability. *Journal of Consulting and Clinical Psychology*, 1971, 36, 338-343.
- Manno, B., & Marston, A. R. Weight reduction as a function of negative covert reinforcement (satiation) versus positive covert reinforcement. *Behavior Research and Therapy*, 1972, 10, 201-207.
- Marston, M. Compliance with medical regimens as a form of risk taking in patients with myocardial

- infarctions. (Doctoral dissertation, Boston University, 1969) *Dissertation Abstracts International*, 1969, 30, 2151A-2152A. (University Microfilms No 69-18,722)
- Martin, R. D., & Shepel, L. F. Locus of control and discrimination ability with lay counselors. *Journal of Consulting and Clinical Psychology*, 1974, 42, 741.
- Mechanic, D. Social psychological factors affecting the presentation of bodily complaints. *New England Journal of Medicine*, 1972, 286, 1132-1139.
- Melges, F. T., & Weisz, A. E. The personal future and suicidal ideation. *Journal of Nervous and Mental Disease*, 1971, 153, 244-250.
- Mlott, S. R., & Mlott, Y. D. Dogmatism and locus of control in individuals who smoke, stopped smoking, and never smoked. *Journal of Community Psychology*, 1975, 3, 53-57.
- Morley, E. L., & Watkins, J. T. Locus of control and effectiveness of two rational-emotive therapy styles. *Rational Living*, 1974, 9, 22-24.
- Moyal, B. R. Locus of control, self-esteem, stimulus appraisal, and depressive symptoms in children. *Journal of Consulting and Clinical Psychology*, 1977, 45, 951-952.
- Moser, A. J. Structural group interaction: A psychotherapeutic technique for modifying locus of control. *Journal of Contemporary Psychology*, 1975, 43, 96.
- Naditch, M. P. Locus of control, relative discontent, and hypertension. *Social Psychiatry*, 1974, 9, 111-117.
- Naditch, M. P. Locus of control and drinking behavior in a sample of men in Army basic training. *Journal of Consulting and Clinical Psychology*, 1975, 43, 96.
- Naditch, M. P., Gargan, M., & Michael, L. B. Denial, anxiety locus of control, and the discrepancy between aspirations and achievement as components of depression. *Journal of Abnormal Psychology*, 1975, 84, 1-9.
- Nowicki, S., & Barnes, J. Effects of a structured camp experience on locus of control in children. *Journal of Genetic Psychology*, 1973, 122, 247-252.
- Nowicki, S., Bonner, J., & Feather, B. Effects of locus of control and differential analogue interview procedures on the perceived therapeutic relationship. *Journal of Consulting and Clinical Psychology*, 1972, 38, 434-438.
- Nowicki, S., & Hopper, A. Locus of control correlated in an alcoholic population. *Journal of Consulting and Clinical Psychology*, 1974, 42, 735.
- Nowicki, S., & Strickland, B. R. A locus of control scale for children. *Journal of Consulting and Clinical Psychology*, 1973, 40, 148-154.
- Obitz, F. W., & Swanson, M. K. Control orientation in women alcoholics. *Journal of Studies in Alcohol*, 1976, 3, 694-697.
- O'Bryan, G. G. The relationship between an individual's IE orientation and information seeking, learning, and use of weight control relevant information (Doctoral dissertation, University of Nevada, 1972). *Dissertation Abstracts International*, 1972, 33, 447B. (University Microfilms No. 72-19,541)
- Olbrisch, M. E. *Perceptions of responsibility for illness and health-related locus of control of gonorrhea patients*. Unpublished master's thesis, Florida State University, 1975.
- O'Leary, M. R., Donovan, D. M., & Hague, W. H. Relationships between locus of control and MMPI scales among alcoholics in replication and extension. *Journal of Clinical Psychology*, 1974, 30, 312-314. (a)
- O'Leary, M. R., Donovan, D. M., & Hague, W. H. Relationship between locus of control, self-report, and nonobtrusive measure of anxiety. *Journal of Clinical Psychology*, 1974, 30, 372-373. (b)
- Oziel, J. R., Obitz, F. W., & Keyson, M. General and specific perceived locus of control in alcoholism. *Psychological Reports*, 1972, 30, 957-958.
- Palmer, R. D. Parental perception and perceived locus of control in psychopathology. *Journal of Personality*, 1971, 3, 420-431.
- Palmore, E., & Luikart, C. Health and social factors related to life satisfaction. *Journal of Health and Social Behavior*, 1972, 13, 68-80.
- Pawlicki, R. E., & Almqvist, C. Authoritarianism, locus of control, and tolerance of ambiguity as reflected in membership and nonmembership in a women's liberation group. *Psychological Reports*, 1973, 32, 1331-1337.
- Phares, E. J. A social learning approach to psychopathology. In J. B. Rotter, J. E. Chance, & E. J. Phares (Eds.), *Applications of a social learning theory of personality*. New York: Holt, Rinehart, & Winston, 1972.
- Phares, E. J. *Locus of control in personality*. Morristown, N.J.: General Learning Press, 1976.
- Phares, E. J., Ritchie, D. E., & Davis, W. L. Internal-external control and reactions to threat. *Journal of Personality and Social Psychology*, 1968, 10, 402-405.
- Phares, E. J., & Wilson, K. G. Responsibility attribution: Role of outcome severity, situational ambiguity, and internal-external control. *Journal of Personality*, 1972, 40, 392-406.
- Phares, E. J., Wilson, K. G., & Klyver, N. W. Internal-external control and the attribution of blame under neutral and distractive conditions. *Journal of Personality and Social Psychology*, 1971, 18, 285-288.
- Pierce, R. M., Schauble, P. G., & Farkas, A. Teaching internalization behavior to clients. *Psychotherapy: Theory, Research, and Practice*, 1970, 7, 217-220.
- Pines, H. A., & Julian, J. W. Effects of task and social demands on locus of control differences in information processing. *Journal of Personality*, 1972, 40, 407-416.
- Platt, J. J., & Eisenman, R. Internal-external control of reinforcement, time perspective, adjustment, and anxiety. *Journal of General Psychology*, 1968, 79, 121-128.
- Powell, A., & Vega, M. Correlates of adult locus of control. *Psychological Reports*, 1972, 30, 455-460.

- Prociuk, T. J., Breen, L. J., & Lussier, R. J. Hopelessness, internal-external control, and depression. *Journal of Clinical Psychology*, 1976, 32, 299-300.
- Pryer, M. W., & Distefano, M. K. Correlates of locus of control among male alcoholics. *Journal of Clinical Psychology*, 1977, 33, 300-303.
- Ray, W. J. The relationship of locus of control, self-report measures, and feedback to the voluntary control of heart rate. *Psychophysiology*, 1974, 11, 527-534.
- Reimanis, G. Effects of locus of reinforcement control and modification procedures in early graders and college students. *Journal of Educational Research*, 1974, 68, 124-127.
- Roth, S., & Bootzin, R. B. The effects of experimentally induced expectancies of external control: An investigation of learned helplessness. *Journal of Personality and Social Psychology*, 1974, 29, 253-264.
- Rotter, J. B. *Social learning and clinical psychology*. Englewood Cliffs, N.J.: Prentice Hall, 1954.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Rotter, J. B. Some problems and misconceptions related to the construct of internal versus external control of reinforcement. *Journal of Consulting and Clinical Psychology*, 1975, 43, 56-67.
- Rotter, J. B., Chance, J. E., & Phares, E. J. *Applications of a social learning theory of personality*. New York: Holt, 1972.
- Sanger, S. P., & Alger, H. A. Dimension of internal-external locus of control and the women's liberation movement. *Journal of Social Issues*, 1972, 28(4), 115-129.
- Schallow, J. R. Locus of control and success at self-modification. *Behavior Therapy*, 1975, 6, 667-671.
- Schlavo, R. S. Locus of control and judgments about another's accident. *Psychological Reports*, 1973, 32, 483-488.
- Seeman, M., & Evans, J. W. Alienation and learning in a hospital setting. *American Sociological Review*, 1962, 27, 772-783.
- Segal, S. M., & DuCette, J. Locus of control and pre-marital high school pregnancy. *Psychological Reports*, 1973, 33, 887-890.
- Seligman, M. E. P. Depression and learned helplessness. In R. J. Friedman & M. M. Katz (Eds.), *The psychology of depression: Contemporary theory and research*. Washington, D.C.: V. H. Winston, 1974.
- Seligman, M. E. P. *Helplessness*. San Francisco, Calif.: Freeman, 1975.
- Shriberg, L. D. Descriptive statistics for two children's social desirability scales, general and test anxiety, and locus of control in elementary school children. *Psychological Reports*, 1974, 34, 863-870.
- Shybut, J. Time perspective, internal versus external control and severity of psychological disturbance. *Journal of Clinical Psychology*, 1968, 24, 312-315.
- Smith, C. E., Pryer, M. W., Distefano, M. K. Internal-external control and severity of emotional impairment among psychiatric patients. *Journal of Clinical Psychology*, 1971, 27, 449-450.
- Smith, R. E. Changes in locus of control as a function of life crisis resolution. *Journal of Abnormal Psychology*, 1970, 75, 328-332.
- Smithyman, S. D., Plant, W. T., & Southern, M. L. Locus of control in two samples of chronic drug abusers. *Psychological Reports*, 1974, 34, 1293-1294.
- Sonstroem, R. J., & Walker, M. I. Relationship of attitude and locus of control to exercise and physical fitness. *Perceptual and Motor Skills*, 1973, 36, 1031-1034.
- Sosis, R. H. Internal-external control and the perception of responsibility of another for an accident. *Journal of Personality and Social Psychology*, 1974, 30, 393-399.
- Steffy, R. A., Meichenbaum, D., & Best, J. A. Aversive and cognitive factors in the modification of smoking behavior. *Behaviour Research and Therapy*, 1970, 8, 115-125.
- Straits, B., & Sechrest, L. Further support of some findings about the characteristics of smokers and nonsmokers. *Journal of Consulting Psychology*, 1963, 27, 282.
- Strassberg, D. S. Relationships among locus of control, anxiety and valued goal expectations. *Journal of Consulting and Clinical Psychology*, 1973, 2, 319.
- Strickland, B. R. The prediction of social action from a dimension of internal-external control. *Journal of Social Psychology*, 1965, 66, 353-358.
- Strickland, B. R. Internal versus external control of reinforcement. In T. Blass (Ed.), *Personality and social behaviors*. Hillsdale, N.J.: Erlbaum, 1977.
- Strickland, B. R. Internal-external expectancies and cardiovascular functioning. In L. C. Perlmutter & R. A. Monty (Eds.), *Choice and perceived locus of control*. Hillsdale, N.J.: Erlbaum, in press.
- Tobias, L. L., & MacDonald, M. L. Internal locus of control and weight loss: An insufficient condition. *Journal of Consulting and Clinical Psychology*, 1977, 45, 647-653.
- Vuchinich, R. E., & Bass, B. A. Social desirability in Rotter's locus of control scale. *Psychological Reports*, 1974, 34, 1125-1126.
- Wagner, C., Bourgeois, A., Levenson, H., & Denton, J. Multidimensional locus of control and voluntary control of GSR. *Perceptual and Motor Skills*, 1974, 39, 1142.
- Wallston, K. A., Maides, S., & Wallston, B. S. Health-related information seeking as a function of health-related locus of control and health value. *Journal of Research in Personality*, 1976, 10, 215-222.
- Wallston, B. S., Wallston, K. A., Kaplan, G. D., & Maides, S. A. Development and validation of the health locus of control scale. *Journal of Consulting and Clinical Psychology*, 1976, 44, 580-585.
- Wareheim, R. G., & Foulds, M. L. Perceived locus

- of control and personal adjustment. *Journal of Consulting and Clinical Psychology*, 1971, 37, 250-252.
- Watson, D. Relationship between locus of control and anxiety. *Journal of Personality and Social Psychology*, 1967, 6, 91-92.
- Weaver, R. *Internality, externality, and compliance as related to chronic home dialysis patients*. Unpublished master's thesis, Emory University, 1972.
- Wendland, C. J. Internal-external control expectancies of institutionalized physically disabled. *Rehabilitation Psychology*, 1973, 20, 180-186.
- Williams, A. F. Factors associated with seat belt use in families. *Journal of Safety Research*, 1972, 4(3), 133-138. (a)
- Williams, A. F. Personality characteristics associated with preventive dental health practices. *Journal of American College of Dentists*, 1972, 39, 225-234. (b)
- Williams, A. F. Personality and other characteristics associated with cigarette smoking among young teenagers. *Journal of Health and Social Behavior*, 1973, 14, 374-380.
- Williams, C. B., & Nickels, J. B. Internal-external control dimension as related to accident and suicide proneness. *Journal of Consulting and Clinical Psychology*, 1969, 33, 485-494.
- Wolk, S. Situational constraint as a moderator of the locus of control adjustment relationship. *Journal of Consulting and Clinical Psychology*, 1976, 44, 420-427.
- Wolk, S., & DuCette, J. Intentional performance and incidental learning as a function of personality and task dimensions. *Journal of Personality and Social Psychology*, 1974, 29, 90-101.
- Wolk, S., & Kurtz, J. Positive adjustment and involvement during aging and expectancy for internal control. *Journal of Consulting and Clinical Psychology*, 1975, 43, 173-178.
- Wortman, C. B., & Brehm, J. W. Responses to uncontrollable outcomes: An integration of reactance theory and the learned helplessness model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 8). New York: Academic Press, 1975.

Received April 8, 1977 ■

Multivariate Classification of Day-Care Patients: Personality as a Dimensional Continuum

Rudie J. Lang

Institute of Psychiatry, University of London
London, England

The categorical versus dimensional view of psychiatric abnormality was examined for three groups of dysthymic neurotics, schizophrenics, and alcoholics. A discriminant function analysis correctly assigned 70% of the patients to their a priori diagnostic profiles, though the "hit rate" rose to 90% when patients were aligned on a continuum of severity in two-dimensional person space. Univariate F ratios showed the three phenotypes to differ on 12 of 26 parameters, yet 18 significant F s emerged when the patients' degree of neuroticism was taken as a putative index of general maladjustment. Results showed that Eysenck's Extraversion and Neuroticism personality factors share a considerable degree of collinearity with the factorially pure Tryon-Stein-Chu trait clusters. Many forms of psychopathology, it is argued, simply reflect gross deviations of the continuously variable dimensions of personality, not discontinuities with qualitative change points. Given a broader context, a "dimensional" model of personality functioning may well, owing to its theoretical quantifiability, supercede the notion of psychiatric "disease" types.

For over 25 years, and perhaps a half-century, the controversy between advocates of the medical model in present-day psychiatry (Panzetta, 1974; Robins, 1976) and its critics, those who prefer to advocate a dimensional system (Eysenck, 1970; Eysenck & Eysenck, 1976; Kendall, 1968; Sjöbring, 1974) has steadily gained momentum. The "dimensional" approach, favored by psychologists, repudiates the notion that functional disorders exist as qualitatively distinct entities, much in the sense that in-

fectious diseases do in *re naturae*. As such, psychiatric theory leads to the practice of viewing a problem as conspicuously present or absent, that is, to labeling a patient either as a hysteric (Briquet's syndrome) or as something else. The underlying assumption on which such terminology and its application rests, inextricably linked to general medicine, is not only of doubtful value but has led to many polemical attacks on the alleged usefulness of the "medical" model in the study and treatment of psychologic morbidity (Eysenck, 1973).

This study was in part based on a Master's of Science thesis submitted by the author to the University of Calgary, Canada.

The valuable support and direction provided by the committee chairman Barry Frost is gratefully acknowledged. Appreciation is also extended to Philip Eaton, Jan Thurgood, and the other staff of the Holy Cross Hospital for their cooperation and support of this research project. However, the views expressed in this article can be attributed solely to the author.

Requests for reprints should be sent to Rudie J. Lang, Institute of Psychiatry, University of London, De Crespigny Park, Denmark Hill, London, SE5 8EA, England.

Instead, the "dimensionalists" claim that a more sensible course to follow is to depict any given debilitating trait (or symptom) on a variable severity continuum, from normal to grossly abnormal. In its simplest terms, this allows the clinician to portray each person according to their infinitely graded *degree* of observable behavior. Obviously, some people, usually in response to real problems and pressures will, at times, appear less their usual extraverted self as they display increasing signs of social isolation, regress from a confident mood to one

of tension or worry, or experience an abrupt cognitive shift from lucid perceptivity to exhibit bizarre delusions of guilt, power, or inferiority. These deviations signal a disruption of the normal continuity of the personality. Any sudden change of emotion, attention, arousal, or perception reflects but one of many intermediate variations on a continuum between healthy and psychiatrically ill states, rather than a clear-cut separation between the two. By analogy, a patient's mental or behavioral state can be portrayed in a manner directly akin to how IQ scores are distributed in the population continuum, fluctuating from severe mental handicap to creative genius.

Briefly, dimensional analysis is based on a number of orthogonal (or independent) personality traits that serve as reference axes, normally distributed and genetically based. Each person is represented by one or more diagnostic points in n -dimensional space, relative to other patients' points. No arbitrary cutoff rules are imposed on the patients' positional loci in multidimensional space. Nor is the dimensional model hampered by the presence of "mixed" or "undiagnosed" psychopathology. Other than pure diagnostic types, the majority of patients will be uniquely placed relative to one or more dimensional reference factors, enabling a reliable probability estimate to be made of the degree to which neurotic or psychotic personality traits are causing the disturbed behavior seen in the patient. It is widely known that in the multivariate domain, *person* space and *variable* space do not correspond to the medical disease entity as it exists in psychiatric theory. As standardization data are accumulated, with broad generalizability, as was assiduously collected for norming intelligence tests, a patient's dimensional profile can guide clinical assessment, choice of treatment modality, and need for community follow-up.

Notably, the major criterion of any disease entity, apart from etiology, course, and outcome, must be one of empirical—not theoretical—discontinuity between nosological types (Kendall, 1975). This, of necessity, implies *mutual exclusivity*.¹ Not only personality disorders, but many other diagnostic

subtypes are "notoriously unreliable categories"; and reviews of studies under optimal conditions of diagnosis (Spitzer & Fleiss, 1974; Walton & Presly, 1973) clearly support this proviso. Similarly, an extensive United Kingdom-United States cross-cultural study has shown that American psychiatrists overdiagnose schizophrenia and underdiagnose affective disorders (Cooper, Kendall, Gurland, Sharpe, & Copeland, 1972). Particularly noteworthy is the lack of satisfactory agreement among clinicians, inconsistent use of terminology, and inherent gaps in the nomenclature itself, further compounded by the indecision whether to rate a trait as pathologic or as a normal variation (Presly & Walton, 1973).

For the present, at least, the psychiatric classification scheme and its newly proposed Diagnostic and Statistical Manual of Mental Disorders (DSM-III), though based on a multiaxial system, is not in a form that permits empirical testing by suitable experiments. Potentially, as every science progresses from descriptive, symptom-based classification to a dimensional framework (Hempel, 1967), the real issue may be more of a question of within what dimensionality the "symptom sign" patterns, or personality traits, lie. Like cognitive abilities, the processes underlying arousal and clinical abnor-

¹ If we single out phenylketonuria and color blindness, they reflect but two mutually exclusive kinds of biological abnormality, each occupying a discrete factor space. Yet both of these disease entities can occupy the same person space in cases in which they co-occur in the same individual. Anxiety and depression, two intermediate conditions often viewed as phenomenologically distinct, not only share the same person space (e.g., as when a patient displays both), but they frequently load on the same factor across studies, confirming their factorial interdependency (Lang & Frost, Note 1). Unlike infectious diseases, surprisingly few of which truly exhibit uniquely identifiable causes, behavioral or personality traits often share a multiple, or overlapping, etiology and, in this sense, cannot be adequately defined by the arbitrary cutoffs imposed by a medically based, nosological psychiatric system. Many other common illnesses such as diabetes mellitus, rubella, peptic ulcer, and coronary heart disease can now be more readily investigated in dimensional than in categorical terms.

malities are often discovered to be complex and continuous rather than discrete. In this sense, they are more likely to be *hierarchically organized within dimensions* and less likely to exist in some either/or paradigm as predicted by psychiatric theory. The typologic model tends to confuse the predisposition with the neuroses and psychoses while ignoring the underlying continuum of mental disturbance.

Since little is actually known about the underlying distribution of clinical pathology in the general population continuum, or about those subclinically predisposed to develop mental conditions, a major task is to specify, as a probability estimate, the likelihood that a person will become behaviorally disturbed. The risk factor will vary and can, in dimensional terms, be expressed as a probability function, such as in the "diathesis model of stress" proposed by Edwards (1969).

Unquestionably, the eventual resolution of the continuum-category debate has major ramifications for the prophylaxis and treatment of the deviant personality and direction dictated to future research trends. If the advent of a "dimensional scientology" is already underway—and not 300 years ahead as Robins (1976) intimates—it must stand or fall on its own merits. In fact, there is good reason to believe that people vary enormously in their predisposition to become stressed, whether genetically or environmentally induced. Physically, anyone can potentially break a leg (i.e., qualitative), yet the person with weak fibrous tissue is much more susceptible to accidental injury (i.e., quantitative). As such distinctions are measurable, a coordinate dimensional view of personality would provide valuable data to assess, in probability terms, the degree to which a referred patient is genetically or socially predisposed to overreact to daily stressors. The dimensional model, owing to its theoretical quantifiability, is obviously superior, as it can be tested empirically.

The continued use of qualitative change points, arbitrarily applied, ensures the kind of diagnostic unreliability that is commonly found. By definition, such points remain undefinable. Perhaps, as Wing (1976) re-

marked: "Some combination of the two . . . may eventually allow clinician and scientist to make the best of both worlds" (p. 394). There is, demonstrably, a considerable degree of collinearity between the dimensional personality model and the psychiatric disease system (Eysenck & Eysenck, 1976) to make this transition possible.

Given some of these empirical and theoretical issues, the main purpose of the study is addressed to three questions. First of all, a discriminant function analysis examined the extent of agreement between psychiatrically assigned diagnoses and computer-generated diagnoses for different subsets of clinical predictor scales. A second purpose was to contrast the categorical-dimensional approaches by taking the dimensionally complex Neuroticism factor as a putative index of severity (or level) of maladjustment to the exclusion of the patients' designated diagnostic codes. The third aim was to check the degree of "collinearity" between two of Eysenck's major personality dimensions, namely, Neuroticism and Extraversion, and behavior as measured by the factorially pure Tryon-Stein-Chu "sign symptom" clusters, derived from the standard Minnesota Multiphasic Personality Inventory (MMPI) item pool. Recent evidence by Wakefield, Yam, Bradley, Doughtie, and Cox (1974) suggests that the Eysenck Personality Inventory and MMPI converge in multivariate space, even though the two instruments represent separate approaches to personality assessment. Much like Eysenck, Royce (1973) pointed out that the many primary factors can be adequately accounted for by three major "superfactors," which Royce calls anxiety, introversion, and superego. As a subsidiary hypothesis, it was expected that the *real* differences in morbidity reside not in the erroneous reliance on a diagnostic label but inherently exist within the configurations of disturbed patients at each symptom level, and presumably of how their scores on critical traits *covary* together.

Method

Subjects

One hundred fifty-seven outpatients recently admitted to the Holy Cross Hospital, Day Care Treat-

ment Centre, located at Calgary, Canada, served as subjects. Fifty-four were male; 103 were female. The mean age of the total group was 37.84 years ($SD = 11.83$ years).

Upon initial admission to the hospital, every patient was first seen by the duty doctor in the emergency clinic and was then interviewed by a senior consultant psychiatrist, who assigned patients to the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-II; American Psychiatric Association, 1968). The classifications were 86—neurotic depression (300.4), 38—alcoholic abuse (303.0), and 33—paranoid schizophrenic (295.3). Clearly, it would have been desirable to have two resident psychiatrists separately assign the working diagnoses, enabling a measure of interrater reliability to be computed. Even so, under optimal conditions with two or more raters, "there are no diagnostic categories for which reliability is uniformly high" (Spitzer & Fleiss, 1974, p. 344). In dealing with such patients, the continued usage of qualitative terms to designate arbitrary cutoff points along a dimensions ensures, almost certainly, the kind of poor reliability that besets psychiatry as a whole. Thus, for all patients who were psychiatrically diagnosed, there is always the unreliability of such diagnoses to contend with. I am well aware of the unreliability of making such diagnostic judgments; however, common practice used in handling the vast majority of new admissions is to have one senior staff psychiatrist assign a differential diagnosis. Therefore, it is quite reasonable to contrast its usefulness against that of the dimensional approach. This article will attempt to clarify this point by showing that the dimensional model is empirically superior, owing to its theoretical quantifiability.

Following the preliminary intake, each patient was interviewed by the day care management team to determine their suitability for the day care program. The psychometric evaluation was carried out in small office-type rooms, usually in two sessions over a 1- to 3-day interval. Most patients were on maintenance drug therapy. The prescribed drug was usually a tricyclic antidepressant (e.g., amitriptyline hydrochloride, imipramine), an anti-anxiety drug (e.g., chlorodiazepoxide, diazepam) or a antipsychotic medication (e.g., pimozide, haloperidol, fluphenazine hydrochloride, trifluoperazine) at moderately low, rather than peak dosage levels, so as not to impair their ability to provide suitable coherent responses.

Alcoholics were detoxified for at least 10–14 days; and no patient who had undergone recent electroconvulsive therapy was tested. Thus, all patients were sufficiently reality oriented to properly understand the nature of the questionnaire forms, and they did not display a marked intellectual deficit.² Due to their deteriorated condition, certain highly disturbed or acutely ill persons, irrespective of clinical diagnosis (those psychotically depressed, in a manic phase, or otherwise badly hallucinated), were simply untestable. Thus few acutely psychotic patients were ipso facto included in the study be-

cause they, behaviorally, precluded themselves. Otherwise, the final sample, excluding these few, formed a fairly representative sampling of psychopathology from the general population continuum.

Measures

Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955). Only the WAIS Vocabulary and Block Design subtests were administered on the basis that in independent analyses of dyads of WAIS subtests, both Maxwell (1957) and Silverstein (1970b) have shown that the Vocabulary–Block Design pair is the "best" combination for computing a prorated full-length WAIS IQ.

Tennessee Self Concept Scale—Clinical and Research Form (TSCS; Fitts, 1965). The TSCS booklet is made up of 100 first-person statements to which the subject responds on a 5-point Likert-type scale, ranging from completely true to completely false. The TSCS contains 90 items that evaluate five external references of self-concept: "physical self," "moral-ethical self," "personal self," "family self," and "social self," supplemented by a 10-item Self-Criticism scale.

Eysenck Personality Inventory (Eysenck & Eysenck, 1968). Form A of this questionnaire was used; it consists of 57 yes–no questions: 24 evaluate neuroticism–emotional stability, 24 evaluate introversion–extroversion, and 9 items constitute the Lie (or "social desirability") scale.

Tryon-Stein-Chu (TSC) scales (Stein, 1968). This 200-item inventory was developed by cluster analyzing the 550 original MMPI item pool. This eliminated the contamination due to item overlap, as only the purer nonoverlap items were retained to reduce the effects of spuriously shared variance. There are eight TSC scales: Social Introversion; Body Symptoms; Suspicion and Mistrust; Depression and Apathy; Resentment and Aggression; Autism and Disruptive Thinking; Tension, Worry, and Fears; and Lying (or Dissimulation).

Nurses' Observation Scale for Inpatient Evaluation (NOSIE; Honigfeld & Klett, 1965). This rating scale contains 80 behavioral items, each of which is rated on a 5-point frequency continuum, ranging from never to always. Only five social adjustment scales were used: Social Competence, Social Interest, Personal Neatness, Cooperation, and Irritability. A team of mental health workers and psychiatric nurses rated each patient on each of these indices following no less than 7 days of careful observation in vivo.

² The mean WAIS IQ, prorated from the summed Vocabulary and Block Design scores using Silverstein's (1970a) conversion formula (age corrected for the 34- to 44-year-old group), was 103 to indicate that patients were of average intellectual ability.

Procedure

Phase 1. To begin, all patients were classified into a priori groups in conformity with their assigned psychiatric diagnoses. A stepwise discriminant function analysis was carried out to ascertain the extent of agreement between the categorical diagnoses and computer-generated profiles for three types of data variables (scores based on observational, self-report, and cognitive assessment). Three combinations of predictor scales were used: (a) all 26 independent variates; (b) the seven TSC subscales; and (c) the Extraversion and Neuroticism factors.

Phase 2. Second, all patients were divided into three clusters along a continuum of neurotic severity. Of the possible choices, Eysenck's Neuroticism was preferred as it represents a universally stable, personality factor. Thus, the 157 cases, regardless of assigned diagnoses, were positioned within a dimensional framework in keeping with their level of maladjustment, namely, Neuroticism (N) scores. The severity continuum consisted of three levels: low N scores (0-7), moderate N scores (8-13) and high N scores (14-23). In this way, it was possible to plot, or align, the patients along a graded, dimensional personality factor and then contrast this approach against the notion of a disjunctive psychiatric entity.

Statistical treatment. Initially, Pearson product-moment correlations were computed for all 26 variates in the raw data matrix. The suitability of the resultant multiple correlation matrix for further multivariate analysis was ensured by the Dzuiban and Shirkey (1974) test, which ruled out the possibility of random fluctuation. Given each subject's observed series of scores, the discriminant program computes a set of lambda weights (the larger lambda is, the less discriminating power present) that are applied to the raw scores of the data variables for each case, producing a standard score for each subject (Cooley & Lohnes, 1971). The individual test scores are then transformed into a single discriminant score, and that score is the patient's locus in n -dimensional space. For two or more groups, the number of discriminant functions is always a potential maximum of $g - 1$ functions, each of which define an orthogonal reference axis in geometric space (Nie, Bent, & Hullz, 1970).

Both the BMD (Dixon, 1972) and FORTRAN (Veldman, 1967) computer programs generate an approximate F test of significance based on Wilk's lambda to determine the equality of group covariance matrices. Lambda can be transformed into a chi-square statistic to test the significance level of each successive function. Likewise, the relative sizes of λ_i (eigenvalues) and their associated canonical correlations denote the ability of each function to maximally separate the groups.

When two discriminant functions are derived, the group centroids (means of each group on all functions) can be plotted in geometric space to give an optimal two-dimensional picture of the degree of group separation. The mathematical logic of dis-

criminant analysis assumes that the data variables are sampled from a multivariate normal distribution having equal variance-covariance matrices within groups. However, owing to the robustness of the technique, these assumptions are often not strongly adhered to. To correct for unequal N sizes, a Bayesian adjustment of the $\log |H|$ sign was made based on a priori knowledge of group membership probabilities. The prior probability values were thus set equal to the cell frequencies. The classifications are based on the separate group covariance matrices, rather than the pooled within-groups covariance (dispersion) matrix, as illustrated by Cooley and Lohnes (1971, chap. 10).

Moreover, the magnitudes of the weighting coefficients, in standardized form, are closely analogous to beta weights in multivariate regression formulas. Hence, like regression coefficients, all discriminant function coefficients change in value with the addition or deletion of a variable from the specified analysis, requiring a recomputation of the functions at each stage. The standardized function coefficients depict the relative contribution of its associated variable to that function. As in factor analysis, these coefficients are often used to "label" each function by identifying the dominant characteristic that they measure. Based on the discriminant score, cases are assigned to the diagnostic category for which the computed probability density is largest. Finally, subprogram DISCRIMINANT tests the significance of the scale means of the measures for each subgroup by univariate and step-down F values.

Results

Analysis 1: Diagnostic Criterion

Twenty-six variables. The test of overall differentiation for the three clinical typologies, Wilk's lambda (the associated chi-square of each discriminant function), and correlations of the 26 predictors are presented in Table 1. For the overall analysis, Wilk's $\Lambda = .652$, which is significant ($p < .001$). The nature of these discriminant functions can be determined by examining the large contributors to group separation. The first function is bipolar, being positively weighted by WAIS Vocabulary; and Neuroticism; Self-Criticism; Introversion; Depression; Tension, Worry, and Fears; Cooperation and Social Interest; and negatively by Extraversion; Personal Self, and the EPI Lie scale. This first function, or ordinate, contains a number of the primaries that comprise Eysenck's Neuroticism or Cattell's Anxiety superfactors.

The second variate, namely the abscissa, has its positive pole defined by Suspicion-

Table 1
Correlations of 26 Variables With Two Discriminant Functions and Tests of Significance for Three Psychiatric Subgroups

Predictor	Discriminant function			
	Diagnostic codes ^a		Symptom severity ^b	
	1	2	1	2
Age	-.19	-.04	-.06	-.19
WAIS Vocabulary	.36	-.07	.16	-.37
WAIS Block Design	-.13	-.08	.04	-.35
Extraversion	-.48	-.44	.21	.45
Neuroticism	.65	.33	—	—
Lie scale (EPI)	-.47	.35	-.05	.41
Physical Self	.15	-.11	.05	-.13
Moral-Ethical Self	-.04	-.31	.77	-.35
Personal Self	-.43	.15	-.49	.14
Family Self	.32	-.03	-.17	-.06
Social Self	.24	.23	-.30	-.11
Self-Criticism	.43	-.08	.67	.18
Social Introversion	.49	.05	.71	.17
Body Symptoms	.11	-.30	.32	.14
Suspicion, Mistrust	-.23	.49	-.30	.78
Depression, Apathy	.69	.23	.63	.19
Resentment, Aggression	.19	.41	-.11	.59
Autistic Thinking	-.11	.57	.26	.61
Tension, Worry, and Fears	.66	.12	.76	.30
Lie scale (MMPI)	.15	.44	-.69	.51
Cooperativeness	.65	.09	.31	-.10
Personal neatness	.51	-.37	.34	.13
Irritability	.08	-.32	.12	.37
Social interest	.38	-.45	-.32	.53
Social competence	.31	-.69	-.14	-.41
Sex gender	.22	-.33	.42	.44
% variance	58.59	41.39	74.43	25.57
Bartlett's χ^2	83.66	35.87	93.19	53.09
df	50	24	49	23
p	.002	.056	.001	.012

Note. WAIS = Wechsler Adult Intelligence Scale; EPI = Eysenck Personality Inventory; MMPI = Minnesota Multiphasic Personality Inventory.

^a Wilk's Λ = .462, $F(2, 154) = 20.90$, $p < .001$.

^b Wilk's Λ = .447, $F(2, 154) = 18.78$, $p < .001$.

Mistrust, the EPI Lie scale, Resentment-Aggression, Autistic Thinking, and the MMPI Lie scale, whereas the negative pole points to deficits in Extraversion, Personal Neatness, Social Interest, and Social Competence. Generally, then, the subset of variates that define this latter function reflect the more acute symptomology typical of the psychoticlike profile. It presumably refers to a Psychoticism dimension.

To facilitate interpretation, the group centroids are plotted in Figure 1. As is apparent, there is substantial overlap between the three typologies. They are not clearly truncated,

even though the discrimination is statistically significant. For the dysthymic neurotics, the total battery classified them with 75.6% efficiency, misclassifying 21 of these 86 patients. Table 2 shows how this ratio dropped to 60.6% agreement between diagnosis and objective tests for schizophrenics, whereas a 65.8% concordance was achieved for the alcoholic group. For the most part, the 30% misclassification rate may be due largely to the speculative nature of the defining criterion, not the standardized objective tests. That the differential diagnoses of 47 of 157 patients showed a poor fit to the clinical data

Table 2
Diagnostic Classification by Discriminant Function Analysis for Different Combinations of Predictor Equations

Measure	Actual group	Predicted group membership		
		1	2	3
26 variates ^a	Neurotics	65 (75.6)	6 (7.0)	15 (17.4)
	Schizophrenics	10 (30.3)	20 (60.6)	3 (9.1)
	Alcoholics	10 (26.3)	3 (7.9)	25 (65.8)
7 MMPI scales ^b	Neurotics	66 (76.8)	7 (8.1)	13 (15.1)
	Schizophrenics	14 (42.4)	17 (51.5)	2 (6.1)
	Alcoholics	12 (31.6)	2 (5.3)	24 (63.1)
2 EPI scales ^c	Neurotics	66 (76.8)	8 (9.3)	12 (13.9)
	Schizophrenics	13 (39.4)	15 (45.4)	5 (15.2)
	Alcoholics	11 (29.0)	4 (10.5)	23 (60.5)

Note. Numbers in parentheses are percentages. For neurotics, $n = 86$; schizophrenics, $n = 33$; alcoholics, $n = 38$. MMPI = Minnesota Multiphasic Personality Inventory; EPI = Eysenck Personality Inventory.

^a $\chi^2(4) = 134.27, p < .001$; 70.70% of cases were correctly identified.

^b $\chi^2(4) = 97.20, p < .001$; 68.15% of cases were correctly identified.

^c $\chi^2(4) = 89.77, p < .001$; 66.24% of cases were correctly identified.

is in accord with their widespread unreliability (Kendall, 1975). Moreover, the test(s) may be more valid than the criterion!

MMPI scales. The overall test of between-group differences, Wilk's lambda criterion, converted to Rao's F approximation, for the abbreviated MMPI (TSC trait clusters) is

highly significant ($p < .001$). Of seven TSC scales, five contributed significantly to group separation. According to Table 3, the two best predictors were Social Introversion and Tension, Worry, and Fears. Despite two significant roots, the discriminant analysis was only able to correctly identify 68.15% of patients as members of the consensual class that they were intended to depict (see Table 2). Classification figures for alcoholics (68.1%) and paranoid schizophrenics (51.5%) were uniformly poor but were somewhat better for dysthymic neurotics (76.8%). As the shorter MMPI (i.e., Midi-Mult) enjoys wide usage as a diagnostic tool, the substantial overlap runs contrary to the notion of a qualitatively distinct, categorical entity.

That a sizable number of patients (50 of 157) do not exhibit finite, qualitative change points in multiperson space provides, at least, a justifiable *raison d'être* for exploring the merits of a dimensional system of personality description. Here again, the absence of any marked points of "rarity" runs counter to classical theory.

EPI scales. Third, a prediction model using only a single pair, the extraversion and neuroticism personality traits, was used. Wilk's lambda for the three-group differentiation was .572, which was significant ($p <$

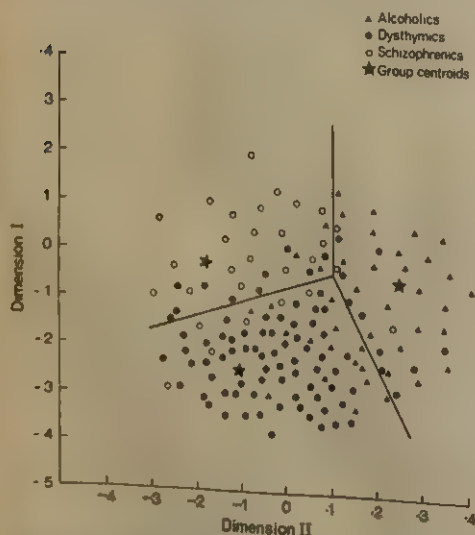


Figure 1. Position of 86 dysthymic neurotics, 38 alcoholics, and 33 schizophrenics on two canonical variates based on 26 discriminant scores. (Latent roots: $\lambda_1 = .461$ (58.6%), $p < .002$; $\lambda_2 = .717$ (41.4%), $p < .056$.)

.001). Two latent roots, respectively, accounted for 85.8% and 14.2% of trace. Given this dyad, the stepwise procedure allocated 66.24% of all cases to their intended conceptual categories. Indeed, this is actually only 2% fewer than the foregoing TSC scales and scarcely 4% less than the overall battery.

For this subanalysis, nearly one half of the alcoholics and about 40% of schizophrenics were erroneously assigned, much like the shorter MMPI. By way of comparison, the 2% (or three patients) higher "hit rate" of the pure TSC validity scales, over the much briefer EPI, is hardly a noteworthy advantage for an inventory directly enquiring about "signs and symptoms." Interestingly, each inventory classified 66 or 86 neurotic depressives (a difficult category). Whenever the aim is to classify patients into psychiatric clusters, certain *scalar combinations* may prove more useful than any single unidimensional scale. Inclusion of too large a number of variables may obscure discrimination just as easily.

Considering the recent evidence by Wakefield et al. (1974) that the primary MMPI subscales may be fully accounted for by Eysenck's dimensional personality model, it

Table 3
Correlations of the MMPI-Derived TSC Scales With Two Discriminant Functions and Tests of Significance for Three Subgroups of Day-Care Patients

Scale	Discriminant function		Univariate <i>F</i>	<i>p</i>
	1 ^a	2 ^b		
Social Introversion	.51	-.76	22.19	.001
Body Symptoms	.31	.14	.98	.084
Suspicion, Mistrust	-.81	.50	8.82	.009
Depression, Apathy	.67	.29	3.95	.019
Resentment, Aggression	.23	.49	1.16	.162
Autistic Thinking	.29	.52	7.93	.010
Tension, Worry, and Fears	.81	.09	9.81	.003
Lie (Dissimulation)	.10	.71	4.55	.019

Note. Wilks' $\Lambda = .589$, $F(2, 154) = 24.18$, $p < .001$. MMPI = Minnesota Multiphasic Personality Inventory. TSC = Tryon-Stein-Chu.
^a $\chi^2(14) = 50.65$, $p < .001$; % of variance = 71.6.
^b $\chi^2(6) = 31.66$, $p < .001$; % of variance = 29.4.

Table 4

Intercorrelation Between MMPI-Derived TSC Scales and Scores on the Eysenck Personality Inventory

Scale	Extra-version	Neuroticism	Lie
Social Introversion	-.58**	.49**	-.10
Body Symptoms	-.27*	.50**	-.11
Suspicion, Mistrust	.03	.44**	-.21*
Depression, Apathy	-.39**	.61**	-.26*
Resentment, Aggression	-.14	.67**	-.37**
Autistic Thinking	-.17	.61**	-.24*
Tension, Worry, and Fears	-.28*	.70**	-.28*
Lying (Dissimulation)	.04	-.22*	.47**

Note. $N = 157$. MMPI = Minnesota Multiphasic Personality Inventory. TSC = Tryon-Stein-Chu.

* $p < .05$ (two-tailed).

** $p < .01$ (two-tailed).

is of interest to note the intercorrelations shown in Table 4. Evidently, the second-order Neuroticism factor is strongly correlated with the factorially pure MMPI validity scales. Had these been corrected for attenuation, the multiple correlation values would have been substantially higher. This moderately high relationship, in terms of shared variance, between the two inventories partially explains why the comparatively short EPI suffers little or no loss in overall predictive power (68.15% vs. 66.24%). The results reaffirm that a good deal of "collinearity" exists between Eysenck's personality traits and psychiatric disorders, as measured by symptom-sign questionnaires such as the MMPI—sufficiently so, it would appear, to suggest that the emphasis on dimensional over categorical criteria is not misplaced.

Analysis 2: Continuum of Severity

Twenty-five variables. The measure of intergroup dispersion, Wilk's lambda, as shown in Table 1, was significant ($p < .001$). Two roots emerged to account for 75.5% and 24.5% of the between-group variance, respectively. The first function had moderately high weightings on Moral Self; Self-Criticism; Social Introversion; Depression; Tension, Worry, and Fears; and the MMPI Lie scale.

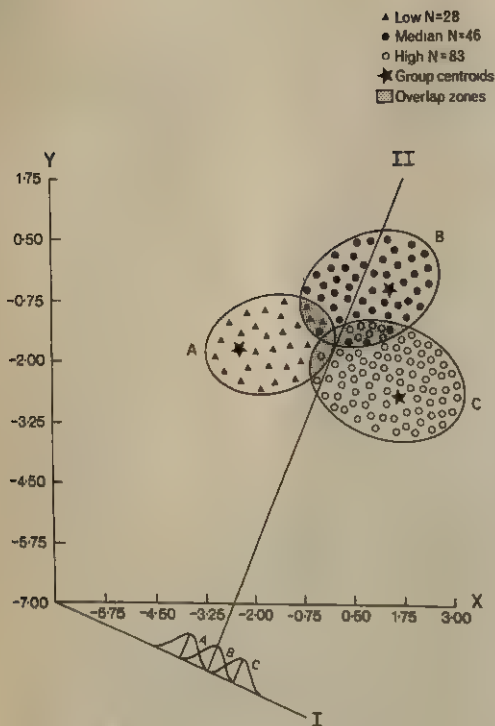


Figure 2. Dimensional profile of 28 low, 46 median, and 83 high-level Neuroticism (N) patients based on 25 predictor variables. (Latent roots: $\lambda_1 = .448$ (75.5%), $p < .001$; $\lambda_2 = .764$ (24.5%), $p < .016$.)

This combination is somewhat suggestive of the intermediate forms of behavioral disturbance usually attributed to the dysthymic personality (i.e., those patients whose problems focus on phobic, anxious, depressive, or obsessive concerns).

The second function was positively loaded by Suspicion-Mistrust, Autistic Thinking, Resentment-Aggression, the MMPI Lie scale, and Extraversion, and negatively by Social Interest. This configuration seems to allude to the distorted reality contact, social deviance, and depersonalization commonly found at the higher endpoint of the normality-psychoticism spectrum. Such vectors may well signal the diffuse, yet quantitative, level at which extreme forms of normal thinking and feeling processes fade into gross psychologic abnormality. If this is indeed the case, it is expected that the more highly disturbed patients would occupy this region of the severity continuum. On any such continuum would lie

the commonly referred to, though vague multiple diagnostic types like pseudoneurotic schizophrenia, schizophreniform personality, chronic anxiety depression, or alcoholic psychosis-paranoid type. These could well be dropped as they simply add to the internal contradictions of categorical psychiatry.

An inspection of Table 5 indicates that for the diagnostic model, only 12 of 26 clinical predictor scales significantly distinguished between the three conceptual typologies. However, for the "degree" of severity continuum, 18 of 25 (excluding Neuroticism itself) predictor variables register a significant level of discriminant validity. Two TSC scales (Tension, Worry, Fears and Resentment-Aggression) were the most powerful predictors for the dimensional analysis. The dimensional approach, even in the present limited context, yields more encouraging results, suggesting that patients can more easily be aligned on a continuous maladjustment axis. Conversely, the search for "pure" or sharply demarcated subtypes not only leads to the misapplication of a science but encourages widespread criticism.

Of considerable interest are the individual points in multiperson space seen in Figure 2. Of the three ellipsoids, one contains all but 3 low Neuroticism cases. The second region contains, with 4 exceptions, the 46 patients divided at their median Neuroticism levels. The third zone is occupied by, with the exception of 7 moderately disturbed patients positioned in the second region and 1 stray subject assigned to the lower end of the continuum, all of the most debilitated patients with highly elevated Neuroticism profiles. There is an impressive degree of separation along the unbounded severity continuum ($p < .001$ by Fisher's exact test) for all three levels of general maladjustment.

In terms of classification efficacy, Table 6 shows that the dimensional model is able to portray the patients' loci in n -dimensional space with 90.44% accuracy taking, as a putative index, each subject's level of Neuroticism. Dimensional analysis, it seems, brings out the "real" covariations with sharper clarity. Naturally, this illustrative procedure remains to be carried out with other carefully

Table 5
Univariate Analysis of Group Means for Diagnostic Versus Dimensional Criteria on 25 Objective Tests

Scale	A priori diagnosis			F^a	Neuroticism severity continuum			F^a
	Neurotics ($n = 86$)	Schizo- phrenics ($n = 33$)	Alcoholics ($n = 38$)		Low ($n = 28$)	Moderate ($n = 46$)	High ($n = 83$)	
Age	36.87	38.86	37.78	.77	36.00	42.57	34.96	1.99
WAIS Vocabulary	51.01	51.62	52.11	.34	55.00	59.43	50.62	3.80*
WAIS Block Design	32.93	31.08	35.22	1.48	35.12	31.73	32.39	.99
Extraversion	9.79	10.00	13.11	3.07*	14.56	10.21	8.15	5.71**
Neuroticism	18.39	14.07	17.72	3.96*	—	—	—	—
Lie (EPI)	2.70	4.31	3.20	3.41*	4.45	3.43	2.33	5.65**
Physical Self	54.58	62.08	55.45	3.93*	64.67	57.07	50.36	6.92**
Moral-Ethical Self	60.12	67.39	59.79	3.16*	59.67	66.86	60.76	2.72
Personal Self	49.02	61.23	48.78	9.25***	57.33	53.62	48.11	7.54**
Family Self	57.45	59.63	54.69	2.01	60.66	57.93	53.20	5.46**
Social Self	56.57	60.62	55.06	1.29	60.33	58.36	53.54	4.75**
Self-Criticism	35.92	33.23	37.68	1.83	32.00	35.07	39.75	5.83**
Social Introversion	18.01	12.31	16.16	4.36*	12.03	14.70	19.73	8.32***
Body Symptoms	17.81	15.07	16.29	1.63	11.67	15.93	21.58	8.69***
Suspicion, Mistrust	14.53	19.07	11.39	5.82**	11.18	14.43	18.92	7.18**
Depression, Apathy	20.68	15.76	20.50	4.42*	15.67	17.07	24.20	14.47***
Resentment,								
Aggression	12.49	10.20	12.53	1.05	8.00	10.37	16.85	20.91***
Autistic Thinking	14.03	19.38	12.72	4.52*	12.44	13.67	20.03	11.98***
Tension, Worry, and								
Fears	23.77	22.84	17.95	4.19*	15.33	20.57	28.69	22.83***
Lie (MMPI)	2.95	4.09	3.28	1.08	5.07	2.79	2.46	8.54***
Cooperation	10.28	8.69	8.72	2.51	10.93	8.72	8.05	2.73
Personal neatness	13.19	11.08	11.56	2.86	12.57	12.64	10.58	1.32
Irritability	6.99	9.23	8.06	1.76	6.01	8.20	10.08	3.54*
Social interest	32.66	25.69	30.72	4.04*	33.44	30.21	27.41	4.01*
Social competence	43.81	38.53	41.89	2.42	43.29	41.29	39.66	1.06
Sex gender	.59	.46	.61	1.16	.49	.54	.62	.77

Note. Patients were divided into low-, moderate-, and high-Neuroticism clusters irrespective of diagnostic profiles. WAIS = Wechsler Adult Intelligence Scale; EPI = Eysenck Personality Inventory; MMPI = Minnesota Multiphasic Personality Inventory.

^a $df = 2, 154$.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

chosen factors for even broader discriminability. If actually more realistic, the dimensional view of personality would, obviously, have to be augmented by sensitive tests of physiological, biochemical, and cognitive functioning.

MMPI scales. The seven TSC scales were combined into a prediction formula that produced two canonical variates; both latent roots were highly significant. Dimension 1 accounted for 64.3% of the variable ($\lambda = .666$), $\chi^2(14) = 50.76$, $p < .001$; whereas the second function accounted for the residual

35.7% of trace ($\lambda = .793$), $\chi^2(6) = 25.42$, $p < .001$. The group centroids (or mean discriminant scores) are illustrated in Figure 3. As previously shown, taking each patient's degree of neurotic variability results in an optimal *interdimensional* (as opposed to *intergroup*) dispersion to pinpoint the locus of 85.35% of all psychiatric patients compared to the 68.15% correct consensus achieved for the diagnostic criterion. Were it not for the relatively high proportion of dysthymic neurotics positively identified, the percentage range would surely have fallen into the mid-

Table 6
Dimensional Dispersion of Patients on a Neuroticism Severity Continuum for Global Versus Tryon-Stein-Chu Diagnostic Scales

Measure	Actual Neuroticism group	Predicted group membership		
		1	2	3
26 variates ^a	Low	25 (89.3)	3 (10.7)	0 (0.0)
	Moderate	1 (2.2)	42 (91.3)	3 (6.5)
	High	1 (1.2)	7 (8.4)	75 (90.4)
7 MMPI scales ^b	Low	24 (85.7)	3 (10.7)	1 (3.6)
	Moderate	3 (6.5)	37 (80.4)	6 (13.0)
	High	2 (2.4)	8 (9.6)	73 (88.0)

Note. Numbers in parentheses are percentages. For Low Neuroticism, $n = 28$; for moderate Neuroticism, $n = 46$, for high Neuroticism, $n = 83$. MMPI = Minnesota Multiphasic Personality Inventory.
^a $\chi^2(4) = 182.79, p < .001$; 90.44% of cases were correctly identified.
^b $\chi^2(4) = 119.13, p < .001$; 85.35% of cases were correctly identified.

dle 50s or 60s in line with those shown for the alcoholic and schizophrenic groups.
Across diagnoses, then, the number of cor-

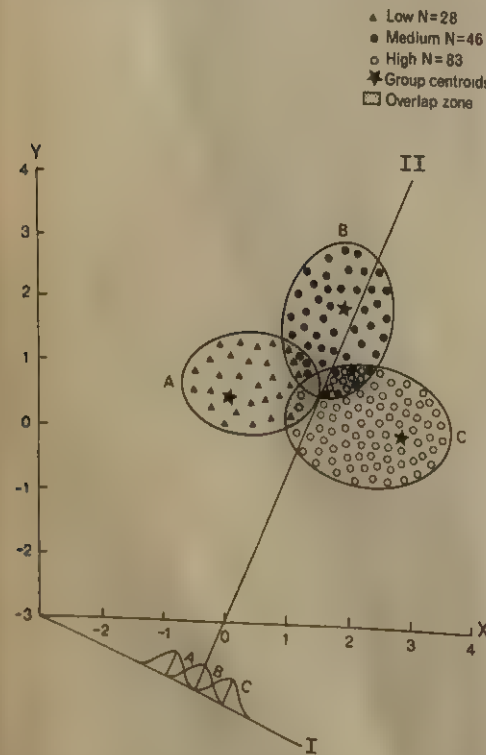


Figure 3. Dimensional position of 157 psychiatric patients on a Neuroticism (N) severity continuum predicted from seven Tryon-Stein-Chu scales. (Latent roots: $\lambda_1 = .666$ (64.3%), $p < .001$; $\lambda_2 = .793$ (35.7%), $p < .001$.)

rectly identified cases for the global dimensional analysis, and seven-variable TSC battery, is clinically superior to that of psychiatric discriminability. Noticeably, the less impressive evidence for the reliability of the diagnostic concepts, as compared with objective laboratory tests, is based on only a conservative number of hospital patients and, as such, needs to be replicated on larger sample sizes, related indices of psychopathology, and alternate experimental designs.
A measure of the observed relations between the assigned psychiatric diagnoses and computer classification matrix was also determined (see Table 7). The Kappa coefficient can be described as an interclass correlation that gives the proportion of diagnostic agreement corrected for chance (Fleiss & Cohen, 1973). For the 26-variable analysis, weighted $K = .500$, rising to .843 for the dimensional approach. Clearly, Kappa for the dimensional analysis is significantly larger ($p < .01$) than Kappa for the diagnostic classification matrices. Notably the weighted Kappa values for the DSM-II diagnostic codes are only slightly higher than those reported by Fleiss, Spitzer, Cohen, and Endicott (1972) for similar psychiatric categories. The magnitude of these values is generally in the lower range expected for well-trained psychiatric personnel. The implication for the clinician is that it may be easier to depict patients in terms of their degree of psychopathology along a num-

ber of continuously variable dimensions than to fit patients into predetermined classes.

Discussion

Unlike studies that have sought to affirm that clinical diagnostic profiles are etiologically distinct, the present study offers some provisional evidence that aligning psychiatric patients along a severity continuum, namely, "degree" of Neuroticism is somewhat superior to allocating them into diagnostic categories. This may possibly apply to Eysenck's Psychoticism factor as well. Such a distinction, doubtless, rests on the extent to which conceptually derived stereotypes coincide with the clinician's experience of the observed characteristics of actual patients. In many instances, T data (objective tests) and L data (life behavior, rated) are used to verify the diagnostic nomenclature, yet they report only nominal agreement. Considering the minor tolerances of empirical methods due to peculiarities of observation, methods of data analysis, and varying sample characteristics, there seems a limit to what even a perfect diagnostic test could do when compared against the psychiatric criterion used. The fairly poor agreement, as reflected in the weighted Kappa values shown in Table 7, raises the issue of whether the broad typologies have any clinical utility beyond that evident for a dimensional model of personality functioning.

In terms of clinical utility, the general equivalence of the EPI to a factorially pure MMPI is rather encouraging, even though the former was not constructed specifically with a view to replicating psychiatric diagnoses. This congruence is reassuring, since both instruments denote separate approaches to personality assessment that converge at the multivariate level. Owing to the larger number of critical distinctions shown in Table 5, it appears that the decision rule to use a patient's degree of emotional instability on a continuum of regression from normality "teases" out some of the real differences much more sharply, and with less obscurity, than a categorical reference system. The two discriminant functions, in com-

Table 7

Overall Agreement (Mean Weighted Kappa) Between Diagnostic Versus Dimensional Profiles Generated by Discriminant Function Analysis for Three Psychiatric Subgroups

Scalar combination	% correctly assigned	M reliability (weighted Kappa)
Diagnostic profile		
26 variates	70.70	.500
7 MMPI scales	68.15	.455
2 EPI scales	66.24	.426
Severity continuum		
25 variates	90.44	.843
7 MMPI scales	85.35	.759

Note. The Neuroticism factor itself served as the putative criterion, leaving only 25 predictors for the global dimensional analysis. MMPI = Minnesota Multiphasic Personality Inventory.

bined form, capture 90.44% of the interdimensional continuity of this major personality factor.

Abolition of diagnostic categories, it is often argued, involves loss of clinical-descriptive, social, and etiological data, including the predictions about outcome and treatment response implied by the original diagnosis. Yet, critics of the dimensional model ignore the fact that *modal profiles* explain lawful covariations between normal and psychologically ill persons in a more meaningful way. All psychiatric diagnoses have their own "degree of certainty," which varies according to whether the etiology is fragmentary or well defined. As few psychiatric phenotypes are immutable, mutually exclusive, or nonreflexive, the art of diagnosis remains a fallible decision process founded on the assumption of *homogeneity* of psychiatric disease types—a still largely unproven axiom. Evidence now emerging from the "spectral analyses" of alcoholic, depressive, and antisocial patient groups strongly suggests, in many cases, that such phenotypes are plainly *heterogenous*, with fairly unknown etiology and neurophysiologic basis.

Notably, assigning any person into a predetermined class, apart from the harmful

effect of an erroneous diagnosis (Morrison & Flanagan, 1978), presupposes much loss of idiographic data, for few labels reflect all the content. Whatever dimensional model is eventually chosen would need to compensate for this loss by providing specific details on the interdimensional axes relevant to the patient's varied symptomology. The patient's prodromal signs would serve to guide clinicians in their choice of the "number of continua needed" problem.

Some recent research has already focused on the extent to which variable hierarchies of personal illness do exist (Foulds & Bedford, 1975) and the degree to which symptom patterns can be fitted within a broader dimensional personality system (Eysenck & Eysenck, 1976). Further study of personality functioning within a combined *hierarchical-dimensional* context may well take us on the path to a scientific realism. This, perhaps, is one of the foremost tasks awaiting present-day behavioral scientists. Paradoxically, early psychiatric theory, has had to undergo a good deal of revisional thinking, as it found itself faced with a large number of overlapping intermediate conditions that did not fit the classical concept, particularly the syndromes. In their everyday experience, clinicians hear far too rich and varied material to fit into discrete or bipolar categories. The lack of strict diagnostic criteria by which to codify the abnormal aspects of human behavior suggests that the pretension to a scientific psychiatry is weakly defensible. Under more careful scrutiny, it is usually only the pure cases—like the Cotard or Capgras syndromes—to which the patient has been assigned, Procrusteslike, that are uniquely categorical. And any such unicorns are rare indeed!

In the personality domain, certain higher order factors such as Neuroticism (otherwise known as anxiety or general emotionality as some prefer) occur in the general population continuum with predictable frequency. The predisposition of being affected will vary according to the degree of stress that each individual can tolerate. A. E. Maxwell (1971) showed that neurotics, affective psychotics, and schizophrenics all exhibited a basic core of symptoms of the type generally

referred to as neurotic. Likewise, Vaz Serra and Pollitt (1975) found that the level of Neuroticism for 100 depressives increased proportionately with their level of depression (as measured by the Beck Depression Inventory), whereas the reverse was true for Extraversion. This is in perfect accord with the present findings.

More likely than not, the plotting of *dimensional profiles* can be effectively applied to the treatment process by providing clinical data on symptom elevation, so as to schedule patients into optimal treatment regimes. Patients could simply be referred to a specific treatment modality depending on the relative position that they inhabited within a multidimensional grid. Apart from simply projecting a subject's scores on personality traits alone into a hypergeometrical space, more precise biological variables such as level of sedation threshold, blood gases, skin conductance, electroencephalogram and electrocardiogram anomalies, blood pressure, or extrapyramidal arousal could be included to pinpoint a person's unique dimensional profile. In this way, the less debilitated person would readily be distinguished from the more severely disturbed person. A person's dimensional position, in the hyperspace defined by the major coordinates selected, would thus govern critical treatment decisions about drug dosage levels, degree of psychopathology, nature of neurological impairment, or ability to benefit from the various forms of therapy. The patient's degree of reality contact, self-assertion, or capacity for love, as assessed by objective laboratory tests and systematic observation techniques, would help determine the client's response to treatment. As this growing body of knowledge evolves from hypothesis to theory to empiricism to culminate as scientific law, the advantages of an applied dimensional approach can be more formally addressed.

Extending the argument, it has been shown that schizophrenics with high levels of autonomic arousal (high sedation threshold) exhibit more paranoid features, whereas those with low levels (low sedation threshold) tend to be retarded, affectively flattened, and socially withdrawn (Claridge, 1972). More

importantly, Claridge emphasized evidence which suggests that certain central nervous system parameters in those predisposed to schizophrenia not only covary together differently in their organizational structure, rather than deviations, but show continuity between normal and pathological thinking. This also applies to normal, neurotic, and psychotic patients as well. Much like personality traits, "many biological phenomena are continuous variables" (Lader, 1975, p. 75). Others, such as Weckowicz, Yonge, Cropley, and Muir (1971), have demonstrated the importance of *level* of arousal and symptom *severity* for choice of treatment, whereas level of neuroticism or psychoticism is a more valid predictor of treatment and outcome of depressive illness than is diagnosis (Kendall, 1969).

Patients who exhibit irrational mental processes or cyclic mood fluctuations (e.g., elation-depression) could readily be delineated on a sympathetic arousal continuum, together with any other relevant features, to distinguish both hypomanic and dysphoric states from normal excitatory states. Aside from simple linear deviation along an "arousal continuum," the *hierarchical organization* of central nervous system processes, as a causal determinant of the atypical personality, can be studied within the context of a dimensional framework. Psychiatric abnormalities can be arranged in a hierarchy (like a Thurstonian taxonomy) to enable a more detailed study of how changes in clinical symptomology *covary* with higher order personality traits. For the most part, many forms of bizarre behavior and psychotic thinking processes appear to exist as continuities of regression from normality, not discontinuities. As such, the deviation from normality is quantifiable, rather than being defined by qualitative change points. Moreover, the dysfunctional personality features can be dimensionally rather than categorically defined, with no loss of vital clinical details. If indeed more realistic, one may wonder why the dimensional view of abnormal personality functioning is not more widely adopted as a suitable replacement for the psychiatric reference system.

As modal profiles, based on dimensional

analysis of normal and psychiatrically ill populations are standardized, with wide generalizability, the clinician will be able to speedily calculate the probability associated with each level (or stage) of symptom severity, either by means of an accessible computer terminal or a handy desk calculator. As data are accumulated for the underlying continuum of clinical pathology, dominant symptoms, mood elevation or depression, and fluctuating delusional states would, of necessity, each have their own set(s) of frequency distributions and tests of significance for the critical parameters under study.

Above all, a purely disease-entity model does not reflect the "best fit" of the conceptual nosology to the *actual* empirical features of the general population continuum. Instead, in its place, a dimensional system of personality functioning, sufficiently modernized, would permit the helping professional to portray psychiatric abnormality in a scientific context, and thereby facilitate evaluation and treatment. To this end, daily clinical judgments that affect treatment strategies and therapeutic outcome would thus be made with an impressive degree of certainty. Given a broader scope, the development of applied dimensional models of behavior and personality remains an intriguing research topic.

Reference Note

1. Lang, R. J., & Frost, B. P. *Some personality features of anxiety and depressive neurotics in a day hospital*. Unpublished material, University of Calgary, Alberta, Canada, 1977.

References

- American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (DSM-II)* (2nd ed.). Washington, D.C.: Author, 1968.
- Claridge, G. S. The schizophrenics as nervous types. *British Journal of Psychiatry*, 1972, 121, 1-17.
- Cooley, W. W., & Lohnes, P. R. *Multivariate data analysis*. London: Wiley, 1971.
- Cooper, J. E., Kendall, R. E., Gurland, B. J., Sharpe, L., & Copeland, J. R. M. *Psychiatric diagnosis in New York and London*. London: Oxford University Press, 1972.
- Dixon, W. J. *BMD biomedical computing programs*. Berkeley: University of California Press, 1972.

- Dzuiban, C. D., & Shirkey, E. C. When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin*, 1974, 81, 358-361.
- Edwards, J. H. Familial predisposition in man. *British Medical Bulletin*, 1969, 25, 58-64.
- Eysenck, H. J. The classification of depressive illnesses. *British Journal of Psychiatry*, 1970, 117, 241-250.
- Eysenck, H. J. (Ed.). *Handbook of abnormal psychology* (2nd ed.). London: Pitman, 1973.
- Eysenck, H. J., & Eysenck, S. B. G. *Manual for the Eysenck Personality Inventory*. San Diego, Calif.: Educational and Industrial Testing Service, 1968.
- Eysenck, H. J., & Eysenck, S. B. G. *Psychoticism as a dimension of personality*. London: Hodder & Stoughton, 1976.
- Fitts, W. H. *Manual for the Tennessee Self Concept Scale*. Nashville, Tenn.: Counselor Recording and Tests, 1965.
- Fleiss, J. L., & Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 1973, 33, 613-619.
- Fleiss, J. L., Spitzer, R. L., Cohen, J., & Endicott, J. Three computer diagnosis methods compared. *Archives of General Psychiatry*, 1972, 27, 643-649.
- Foulds, G. A., & Bedford, A. Hierarchy of classes of personal illness. *Psychological Medicine*, 1975, 5, 181-192.
- Hempel, C. G. Introduction to problems in taxonomy. In J. Zubin (Ed.), *Field studies in the mental disorders*. New York: Grune & Stratton, 1967.
- Honigfeld, G., & Klett, J. C. The Nurses' Observation Scale for Inpatient Evaluation. *Journal of Clinical Psychology*, 1965, 21, 65-71.
- Kendall, R. E. *The classification of depressive illness* (Maudsley Monograph No. 18). New York: Oxford University Press, 1968.
- Kendall, R. E. The continuum model of depressive illness. *Proceedings of the Royal Society of Medicine*, 1969, 62, 335-339.
- Kendall, R. E. *The role of diagnosis in psychiatry*. London: Blackwell, 1975.
- Lader, M. H. Arousal measures and the classification of affective disorders. In M. L. Kietzman, S. Sutton, & J. Zubin (Eds.), *Experimental approaches to psychopathology*. New York: Academic Press, 1975.
- Maxwell, A. E. Multivariate statistical methods and classification problems. *British Journal of Psychiatry*, 1971, 119, 121-127.
- Maxwell, E. Validities of abbreviated WAIS scales. *Journal of Consulting Psychology*, 1957, 21, 121-126.
- Morrison, J. R., & Flanagan, T. A. Diagnostic errors in psychiatry. *Comprehensive Psychiatry*, 1978, 19, 109-117.
- Nie, N. H., Bent, D. H., & Hullz, C. H. *Statistical package for the social sciences (SPSS)*. New York: McGraw-Hill, 1970.
- Panzetta, A. F. Toward a scientific psychiatric nosology. *Archives of General Psychiatry*, 1974, 30, 154-171.
- Presly, A. S., & Walton, H. J. Dimensions of abnormal personality. *British Journal of Psychiatry*, 1973, 122, 269-276.
- Robins, E. Categories versus dimensions in psychiatric classification. *Psychiatric Annals*, 1976, 6, 368-374.
- Royce, J. R. The conceptual framework for a multifactor theory of individuality. In J. R. Royce (Ed.), *Multivariate analysis and psychological theory*. London: Academic Press, 1973.
- Silverstein, A. B. Reappraisal of the validity of a short form of Wechsler's scales. *Psychological Reports*, 1970, 26, 559-561. (a)
- Silverstein, A. B. Reappraisal of the validity of WAIS, WISC, and WPPSI short forms. *Journal of Consulting and Clinical Psychology*, 1970, 34, 12-14. (b)
- Sjöbring, H. Mental constitution and mental illness. In S. R. Hirsch & M. Shepherd (Eds.), *Themes and variations in European psychiatry*. Bristol, England: Wright, 1974.
- Spitzer, R. L., & Fleiss, J. L. A re-analysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry*, 1974, 125, 341-347.
- Stein, K. B. The TSC scales: The outcome of a cluster analysis of the 550 MMPI items. In P. Reynolds (Ed.), *Advances in psychological measurement* (Vol. 1). Palo Alto, Calif.: Science & Behavior Books, 1968.
- Vaz Serra, A., & Pollitt, J. The relationship between personality and the symptoms of depressive illness. *British Journal of Psychiatry*, 1975, 127, 211-218.
- Veldman, D. J. *FORTAN programming in the behavioral sciences*. New York: Holt, Rinehart & Winston, 1967.
- Wakefield, J. A., Yam, B. H. L., Bradley, P. E., Doughtie, E. B., & Cox, J. A. Eysenck's personality dimensions: A model for the MMPI. *British Journal of Social and Clinical Psychology*, 1974, 13, 413-420.
- Walton, H. J., & Presly, A. Use of a category system in the diagnosis of abnormal personality. *British Journal of Psychiatry*, 1973, 122, 259-268.
- Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation, 1955.
- Weckowicz, T. E., Yonge, K. A., Cropley, A. J., & Muir, W. Objective therapy prediction in depression: A multivariate approach. *Journal of Clinical Psychology*, 1971, 27, 3-29.
- Wing, J. K. The functional psychoses and neuroses. *Psychiatric Annals*, 1976, 6, 393-403.

Cued Recall and Discrimination of Memory Deficit

P. Rawling

St. Vincent's Hospital, Sydney, Australia

J. G. Lyle

University of Sydney, Australia

Some ambiguities in the scoring procedure for the Logical Memory subtest of the Wechsler Memory Scale were eliminated by adopting a cued recall technique for eliciting responses. A cued recall technique discriminated significantly better than the Logical Memory subtest between memory-impaired and control groups. This appears to have been due to the fact that on Logical Memory, controls showed worse performance on the final sections of the memory passages, whereas this did not occur in cued recall.

The Wechsler Memory Scale (WMS) has an important place in clinical assessment, since despite the restricted size of the normative population and the small variances on some of the subtests, it is one of the few memory tests that is generally available. One of the more important of the WMS subtests is Logical Memory. A number of factor analytic studies have shown that this subtest, together with Associate Learning and Visual Reproduction, load on a factor involving the acquisition and immediate recall of complex stimuli, and that of the seven WMS subtests, these three subtests appear to be the most useful in the detection of cerebral dysfunction (Bachrach & Mintz, 1974; Kear-Colwell, 1973).

In the case of the Logical Memory subtest, however, there are several weaknesses and ambiguities in the administrative and scoring procedures that affect its reliability, validity, and hence its clinical utility. (a) The division of the story into memory units is essentially an arbitrary procedure. (b) The scoring criterion is poorly defined; it is left to the clinician to decide whether synonyms, paraphrasing statements, or approximations should be credited. (c) Many studies sug-

gest that in recalling meaningful prose, subjects will reconstruct a story so as to keep the general theme intact rather than attempt to reproduce it verbatim. Summing the number of words or memory units reproduced verbatim from the original passage may not appropriately assess this style of recall.

This study began as a practical attempt to devise and test a modified version of the Logical Memory subtest that would avoid some of the shortcomings outlined above. A cued recall procedure was devised whereby the subject would be presented orally with a standard list of questions after the reading of each passage. Subjects no longer needed to be uncertain whether they were meant to recall the gist of the passage or attempt to recall it verbatim. In addition, by providing a detailed statement of the scoring criteria for each of the standard questions, the objectivity of scoring such a test should be greatly enhanced.

Subjects (all males) were selected from a local Veterans hospital. Excluded from the study were those who were disorientated or who suffered any motor, linguistic, or psychiatric symptoms likely to interfere with performance on the test battery. Subjects were allocated to the memory deficit and control groups on the basis of their performance on two standardized memory tests, one involving recall of a number of pictures (Simpson Memory Pictures) and the other involving recall of a sequence of meaningless shapes (Simpson Shapes Test). This em-

The authors would like to thank F. N. Simpson, Senior Psychologist, Department of Veterans Affairs (Sydney) for permission to use tests and normative data prior to publication.

Requests for reprints should be sent to Peter Rawling, c/o Caritas Centre, St. Vincent's Hospital, New South Wales 2010, Australia.

pirical method of subject allocation was adopted in recognition of the fact that documented evidence of a cerebral lesion is, of itself, no guarantee of memory loss. All subjects were also administered the Simpson Adult Vocabulary Scale as a control measure of intelligence, since it is known that vocabulary tests are relatively resistant to aging and dementia.

Finally, two groups of subjects were selected. The age range of the subjects was 50–64 years; mean age of the deficit group was 56.73 years ($SD = 3.84$) and 57.40 years for the control group ($SD = 4.03$). The mean IQ of the deficit group was 97.67 ($SD = 11.12$), and for the control group it was 99.37 ($SD = 10.75$).

Both passages of the Logical Memory subtest of the WMS and a modified version of this test involving cued recall by means of standard questions were administered. Two prose passages that comprised cued recall differed in content from those of Logical Memory so that there would be minimal transfer of learning from one to the other. All five tests were administered in a counter-balanced order to avoid order effects; cued recall and Logical Memory were never administered consecutively. Raw scores of Logical Memory and cued recall were converted to *T* scores for direct comparison of means and standard deviations. These scores were subjected to analysis of variance.

Mean scores for Logical Memory were 45.24 for the deficit group ($SD = 8.05$) and 55.15 for the controls ($SD = 8.71$). Mean scores for cued recall were 42.42 for the deficit group ($SD = 6.10$) and 57.61 for the controls ($SD = 6.22$). A split-plot analysis of variance, in which the main effects were groups and tests, was carried out on the standardized scores. The focus of interest in this analysis was the Groups \times Tests interaction, which was significant, $F(1, 58) = 5.48$, $p < .05$. This indicates that cued recall differentiated between the groups significantly better than did Logical Memory.

Further examination of the discriminating value of Logical Memory and cued recall was carried out by comparing group recall scores for the initial, middle, and final thirds of each separately. For Logical Memory, the

two passages were each divided into thirds according to the total number of memory units, and the number of units recalled from each third were then summed across both passages. For cued recall, three questions related to the content of each third of each passage; part scores were the sum of questions correctly answered from each third of both passages.

Split-plot analyses of variance, in which the main effects were groups and parts, were carried out on Logical Memory and cued recall separately. Logical Memory significantly discriminated between the groups, $F(1, 58) = 18.46$, $p < .001$; the parts of the test differed significantly, $F(2, 116) = 13.47$, $p < .001$; and there was a significant Groups \times Parts interaction, $F(2, 116) = 6.08$, $p < .01$. The two groups were discriminated by Logical Memory significantly better in the first and middle thirds of the passages than they were in the final third. Cued recall discriminated between the groups highly significantly, $F(1, 58) = 132.00$, $p < .001$; there was a significant difference between parts of the test, $F(2, 116) = 18.91$, $p < .001$ (with the first third being better recalled than the final third); but the Groups \times Parts interaction was nonsignificant. It was apparent from the analysis that the superior discrimination of cued recall relative to Logical Memory was due to the fact that recall of the controls reduced sharply in the final third of Logical Memory, whereas in cued recall the separation of the groups was maintained throughout the test.

In summary, it was found that a modified version of the Logical Memory subtest of the WMS, in which a cued recall technique was used, discriminated between a memory-impaired group and a control group significantly better than did the standard Logical Memory subtest. This was due to the reduced performance of the controls on the final third of the prose passages of Logical Memory, which may have been the result of proactive inhibition, whereby the recalling of the earlier sections interfered with recall of the final sections.

It was noted qualitatively that three types of errors seemed to characterize the memory-impaired group. (a) *Generalization*: This oc-

occurred when patients responded in the right category but were inaccurate; for example, Smith in lieu of Jackson; 10 o'clock in lieu of 2:30. The controls made 12 such errors, whereas the deficit group made 40 such errors. This finding is contrary to that of Talland (1965), who found that this type of error did not discriminate between impaired and normal groups. (b) *Fabrication*: This occurred when patients introduced material that was not in the original passage. This type of response occurred 8 times in the deficit group and once in the control group. (c) *Contamination*: Here, material from Passage A was given in response to questions pertaining to Passage B. Patients from the

deficit group made 3 such errors, control group patients made no such errors.

References

- Bachrach, H., & Mintz, J. The Wechsler Memory Scale as a tool for detection of mild cerebral dysfunction. *Journal of Clinical Psychology*, 1974, 30, 58-60.
- Kear-Colwell, J. J. The structure of the Wechsler Memory Scale and its relationship to "brain damage." *British Journal of Social and Clinical Psychology*, 1973, 12, 384-392.
- Talland, G. A. *Deranged memory: A psychonomic study of the amnesic syndrome*. New York: Academic Press. 1965.

Received May 18, 1977 ■

Effects of Differences in Suggestibility Within Self- and External-Control Conditions

Gary T. Souheaver and W. John Schuldt
University of Arkansas

This research was designed to study possible effects of differences in waking suggestibility on performance within self- and external-control conditions. Ten high-suggestible subjects and 10 low-suggestible subjects, as measured by body sway, were assigned to each of three experimental conditions—self-control, external-control, and no reward. Response rates of self and external groups were higher than the no-reward group. However, response rates of high-suggestible subjects in the self-control condition were not significantly different than similar subjects in the no-reward group. Moreover, performance of high- and low-suggestible subjects was not significantly different in the external-control condition. *Keywords:* suggestibility, self-control, external-control. High-suggestible subjects in the self-control condition.

Suggestibility, in a hypnotic state and in a normal waking state has long been of clinical interest. Moreover suggestibility has been investigated within a wide range of experimental conditions and populations (e.g., Barber, 1963; Dargatzis & Nassis, 1973; Eysenck, 1967; Jones & Hagen, 1971; Morgan & Hirsch, 1971; Veltman, 1974; Webb, 1975; Weissman, 1973). However, in a review of suggestibility Evans (1967) indicated that the literature on the relationship between suggestibility and conditioning was desultory.

One study that the existence of a relationship between waking suggestibility and response rates in external conditioning is reported by Rasmussen & Hill (1961). They found that high-suggestible subjects, as measured by body sway, responded more frequently to a tone than low-suggestible subjects. This relationship was not observed in a study by Webb (1975).

This study is based on a design suggested by the first author in the first section of the second section of the manuscript. It is a study of the relationship of waking suggestibility and response rates.

Keywords: suggestibility, self-control, external-control, body sway, response rates, waking suggestibility, conditioning.

Weiss, Ullman, & Krasner, 1960). Thus, waking suggestibility may not be related to acquisition of responses, but it may serve as a moderator variable for modification and/or maintenance of existing behavioral patterns.

This view seems congruent with Thoresen and Mahoney's (1974) view that suggestibility may influence self-regulatory strategies. Moreover, it seems consistent with Tolor's (1971) evidence that high-suggestible subjects are externally oriented.

Even though there is evidence that self-reward versus external reward occurring within operant conditioning procedures has little differential effects on performance (Bandura & Perloff, 1967; Liebert, Speigler, & Hall, 1970), there have been no systematic attempts to examine possible effects of differences in waking suggestibility on performance within self- and external-control conditions. Thus, this research was designed to assess the effects of self-control and external control of rewards within groups of subjects who differ in levels of waking suggestibility. Specifically, the following hypotheses were tested: (a) There will be more responding in self- and external-control conditions than in no-reward conditions, (b) high-suggestible subjects will respond more to external control conditions than will low-

suggestible subjects, and (c) low-suggestible subjects will respond more to self-control conditions than will high-suggestible subjects.

Method

Subjects

Sixty volunteer undergraduate students, single and between the ages of 18 and 25, participated as subjects. Restrictions on age and marital status were imposed, since suggestibility has been found to vary as a function of these variables (Morgan & Hilgard, 1973).

Apparatus

The index of suggestibility was postural body sway. An apparatus, similar to that described by Webb (1962a), was designed to convert horizontal body movement into movements of a marker along a vertical scale; that is, a nylon string attached to the subject's collar activated an ink marker positioned behind the subject.

Apparatus for the self-reward, external-reward, and no-reward conditions was similar to that described by Liebert et al. (1970). The subject was seated in front of a panel that contained two parallel columns of lights—20 red lights on the left and 20 green lights on the right. Adjacent to each fifth pair of lights, the numbers 5, 10, 15, and 20 were printed in ascending order. The subject activated the red lights (which served to indicate the criterion for delivery of reward) by turning a criterion selector knob at the center of the panel to one of four positions corresponding to the selected criterion. The green lights were successively lit by turning a hand crank positioned in front of the subject. Two complete clockwise turns activated one green light. When the subject turned the crank enough times to reach the preset criterion, lights on both columns extinguished to signify the end of a trial. A storage container for rewards (tokens) was placed adjacent to the nondominant hand of the subject.

Procedure

Each subject was initially requested to stand, facing a blank wall, on a floor marker 4 feet (1.2 m) from the suggestibility apparatus. After the body sway string was attached to the subject's clothing, he/she was told to close the eyes, relax, and allow both hands to hang loosely at the side with both feet together. The subject was told that he/she would not be shocked, injured, or embarrassed in any manner. The following instructions were presented for 2.5 minutes on a tape recording

played on a standard cassette player:

I want you to listen carefully to what this tape says while you go on just standing there, quite still and relaxed, with your eyes closed. Now, just keep standing there please, quite still and relaxed with your eyes closed, and think of nothing in particular. Just keep standing there quite still and relaxed, and listen to me. Now I want you to imagine that you are falling forward, you are falling, falling forward, falling forward all the time. . . . (adapted from Eysenck, 1947)

Subjects whose body sway exceeded 2 inches (5.1 cm) were considered high suggestible; those who swayed less than 2 inches were considered low suggestible. These operational definitions of suggestibility are consistent with those of Eysenck (1947). Ten high-suggestible subjects and 10 low-suggestible subjects were assigned to each of the three experimental conditions—self-control, external control, and no reward.

Procedures for the self-control and external-control conditions were similar to those of Bandura and Perloff (1967). Subjects in the self-control condition were allowed to select their own reward criterion and to self-administer tokens. These subjects were instructed to choose their performance criterion on the reward apparatus, and whenever the criterion was attained by turning the crank, they were to give themselves a token with the nondominant hand by placing it in the token container. They were also told that these tokens could be exchanged for money, at the rate of one penny per token, at the end of the experiment. Subjects in the self-control condition were told that they could change the performance criterion only once, either higher or lower.

Subjects in the external-control condition did not select a reward criterion. Rather, their criterion was individually yoked with self-control subjects. This allowed for a control of possible reinforcement scheduling effects. External-control subjects were told that they would receive tokens, exchangeable for pennies, each time they turned the crank a sufficient number of times to turn off the red light. Each time the criterion was reached, the experimenter placed a token in the storage container.

Subjects in the no-reward condition were also individually yoked to subjects in the self-control condition. They were told to turn the crank with the dominant hand. Performance was not rewarded with tokens.

Results

Body sway means for high-suggestible subjects [3.58 inches (8.9 cm)] and low-suggestible subjects [1.01 inches (2.5 cm)] were significantly different, $t(58) = 11.74$, $p < .001$. Mean ages of the groups (high

suggestible = 19.10; low suggestible = 19.70) did not differ significantly ($t < 1$). Moreover, mean grade point average of 2.84 and 2.82 for high- and low-suggestible subjects, respectively, did not differ significantly ($t < 1$).

The dependent measure for the three experimental conditions was number of crank turns performed by each subject within 4 minutes (Liebert et al., 1970). Means for each treatment, at both levels of suggestibility, are presented in Table 1.

Since it was predicted that self-control and external-control subjects would respond more than no-reward subjects, regardless of levels of suggestibility, an analysis of variance, collapsing across levels of suggestibility, was conducted. Self-control and external-control subjects manifested significantly greater performance than did no-reward subjects, $F(1, 54) = 14.27, p < .001$.

Dunnett critical differences tests (Keppel, 1973) were used to compare responses of no-reward subjects to subjects within each reward condition at each level of suggestibility. Mean performance of low-suggestible subjects in both the external-control self-control conditions was significantly greater ($p < .05$) than low-suggestible subjects in the no-reward condition. High-suggestible subjects in the external-control condition responded significantly more than did high-suggestible subjects in the no-reward condition ($p < .05$). However, high-suggestible subjects in the self-control condition did not perform significantly more than did high-suggestible subjects receiving no-reward ($p > .05$). Thus, conditioning was not evident for high-suggestible subjects in the self-control condition. It was also predicted that high-suggestible subjects would respond more than low-suggestible subjects in the external-control condition and, conversely, that low-suggestible subjects would respond more than high-suggestible subjects in the self-control condition. High-suggestible subjects did not respond significantly more than low-suggestible subjects in the external-control condition ($t < 1$). However, low-suggestible subjects did respond significantly more than high-suggestible subjects in the self-control

Table 1
Mean Number of Turns

Suggestibility level	Treatment group		
	Self	External	No reward
High	218.20	250.70	172.00
Low	276.40	268.40	126.00

treatment, $t(18) = 2.11, p < .05$. No significant difference was noted between high- and low-suggestible subjects within the no-reward treatment, $t(18) = 1.42, p > .05$.

Discussion

This study demonstrated that increased responsivity resulted with either self-control or external-control procedures when levels of suggestibility were disregarded. However, there was no evidence of conditioning for the high-suggestible subjects in the self-control condition. Thus, self-control procedures may not be equally as effective as external-control procedures when "personality" variables are considered.

It was hypothesized that high-suggestible subjects would respond more to external control than would low-suggestible subjects. This prediction was based on the conceptualization that high-suggestible persons are more susceptible to influence from others and, therefore, are more amenable to external imposition of standards and rewards to increase performance levels. However, the results of this study do not support this prediction; that is, the high- and low-suggestible subjects responded at essentially the same rate in the external-control condition. This finding is not congruent with the findings of Webb (1962b) and Weiss et al. (1960), who reported that high-suggestible subjects did respond more than low-suggestible subjects under response-contingent external reward conditions. These studies, however, used verbal praise as rewards in verbal conditioning procedures. The present study, in contrast, used tokens as rewards in an effortful motor task. Thus, differences in procedure may account for the apparently discrepant results.

Another hypothesis was that low-suggestible subjects would respond more to self-control than would high-suggestible subjects. This prediction was based on the conceptualization that low-suggestible persons are less susceptible to influence from others and are, therefore, less likely to require external imposition of standards and rewards to increase performance rates. The results were consistent with this prediction.

Based on the results of this research, one can speculate that low-suggestible persons respond equally well to either self- or external demands and rewards, whereas high-suggestible persons are so dependent on external factors that their performance is minimal when self-determination of reward contingencies is required. Moreover, self-delivered rewards may not have reinforcement value for high-suggestible persons when they are allowed to determine their own standards.

Our findings, if supported by future research, would seem to have clinical significance. For example, it appears that either external-control or self-control techniques would be effective with low-suggestible clients. Additionally, it would seem that high-suggestible clients would not initially benefit from psychotherapy procedures in which self-control is advocated or imposed. However, caution is urged in applying these preliminary findings to clinical practice, since there is a lack of research relating suggestibility to self-control and because this study may have considerable analogue error.

References

- Bandura, A., & Perloff, B. Relative efficiency of self-monitored and externally imposed reinforcement systems. *Journal of Personality and Social Psychology*, 1967, 7, 111-116.
- Barber, T. X. Physiological effects of "hypnotic suggestions": A critical review of recent research. *Psychological Bulletin*, 1965, 63, 201-222.
- Berreman, J. V., & Hilgard, E. R. The effects of personal heterosuggestion and two forms of auto-suggestion upon postural movement. *Journal of Social Psychology*, 1936, 7, 289-300.
- Das, J. P., & O'Connor, N. Body-sway suggestibility in paranoid and nonparanoid schizophrenics. *International Journal of Clinical and Experimental Hypnosis*, 1959, 7, 121-128.
- DeVoge, J. T., & Sachs, L. B. The modification of hypnotic susceptibility through imitative behavior. *International Journal of Clinical and Experimental Hypnosis*, 1973, 21, 70-77.
- Evans, F. J. Suggestibility in the normal waking state. *Psychological Bulletin*, 1967, 67, 114-129.
- Eysenck, H. J. *Dimensions of personality*. London: Routledge & Kegan Paul, 1947.
- Foreyt, J. P., & Hagen, R. L. Covert sensitization: Conditioning or suggestion? *Journal of Abnormal Psychology*, 1973, 82, 17-23.
- Keppel, G. *Design and analysis: A researcher's handbook*. Englewood Cliffs, N.J.: Prentice-Hall, 1973.
- Liebert, R. M., Spiegler, M. D., & Hall, M. Effects of the value of contingent self-administered and noncontingent externally imposed reward on children's behavioral productivity. *Psychonomic Science*, 1970, 18, 245-246.
- Morgan, A. H., & Hilgard, E. R. Age differences in susceptibility to hypnosis. *International Journal of Clinical and Experimental Hypnosis*, 1973, 21, 78-85.
- Thoresen, C. E., & Mahoney, M. J. *Behavioral self-control*. New York: Holt, Rinehart, & Winston, 1974.
- Tolor, A. Are the alienated more suggestible? *Journal of Clinical Psychology*, 1971, 27, 441-442.
- Velten, E. A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 1978, 6, 473-482.
- Webb, R. A. The postural sway measurement of suggestibility: A methodological and reliability study. *International Journal of Clinical and Experimental Hypnosis*, 1962, 10, 155-161. (a)
- Webb, R. A. Suggestibility and verbal conditioning. *International Journal of Clinical and Experimental Hypnosis*, 1962, 10, 275-279. (b)
- Weiss, R. L., Ullman, L. P., & Krasner, L. On the relationship between hypnotizability and response to verbal operant conditioning. *Psychological Reports*, 1960, 6, 59-60.
- Wickramasekera, I. Effects of electromyographic feedback on hypnotic susceptibility: More preliminary data. *Journal of Abnormal Psychology*, 1973, 82, 74-77.

Received May 23, 1977 ■

Self-Directed Treatment for Premature Ejaculation

Robert A. Zeiss
University of Oregon

One promising variation of the now standard Masters and Johnson approach to treating premature ejaculation lies in the use of self-administered treatment manuals. In a test of one such manual, couples with premature ejaculation problems were assigned randomly to (a) totally self-administered treatment; (b) self-administered treatment in conjunction with minimal therapist (telephone) contact; or (c) standard therapist-administered treatment. Couples were successfully treated by therapists or by themselves when they maintained minimal contact with a therapist. Couples working on their own, with no therapist contact, failed to complete treatment successfully. Follow-up data indicated that although there was deterioration in therapeutic gain following the termination of treatment, improvement over pretreatment responses was maintained on all relevant measures. An analysis of posttreatment data indicated that greatest improvement in ejaculatory control occurred when couples continued to use the squeeze or pause to delay ejaculation, but significant improvement in latency to ejaculation also occurred when couples used neither technique to lengthen intercourse.

During the past few decades, a number of approaches to the treatment of premature ejaculation have appeared. These have included strategies relying on pharmacologic intervention (e.g., Aycock, 1949; Bennett, 1961; Boneff, 1971; Mellgren, 1967), as well as strategies relying on psychotherapy (e.g., Cooper, 1969) or combining chemotherapy with psychotherapy (e.g., Friedman, 1968; Schapiro, 1943). In the 1950s, Semans (1956) and Wolpe (1954, 1958) introduced learning-oriented approaches to the problem; these approaches were largely ignored until the work of Masters and Johnson (1970) refined the approaches and revived

interest in the behavioral treatment of premature ejaculation.

Since the Masters and Johnson publication, most research in this area has focused on increasing the efficiency of their highly effective approach (e.g., Clarke & Parry, 1973; Kaplan, 1974; Kaplan, Kohl, Pomeroy, Offit, & Hogan, 1974; Zeiss, Christensen, & Levine, 1978; Zilbergeld, 1975). However, few of these studies have presented any data to support contentions of therapeutic effectiveness and efficiency.

Self-Directed Treatments

A recent and promising direction in the treatment of sexual dysfunction has been the development of behavioral self-help programs written to allow couples to treat sexual problems on their own, without the extensive intervention of a professional therapist. These have included programs for the treatment of general sexual dysfunction (e.g., Kass & Strauss, 1975) and female orgasmic dysfunction (e.g., Barbach, 1975; Heiman, LoPiccolo, & LoPiccolo, 1976; Kline-Graber & Graber, 1975), as well as specifically for

This report is based on a doctoral dissertation submitted to the University of Oregon.

The author gratefully acknowledges the assistance of Debra Jackson-Spangler, Robert Kurlychek, Dennis McClure, Terry S. Trepper, and Katie Whalen, all of whom served as therapists in this study. The encouragement and contributions of Antonette Zeiss were invaluable in all stages of the project.

Requests for reprints should be sent to Robert A. Zeiss, who is now at Valle Del Sol, 1209 South First Avenue, Phoenix, Arizona 85003.

premature ejaculation (Lowe & Mikulas, 1975; Vandervoort & McIlvenna, 1972; Zeiss & Zeiss, 1978). The treatment for premature ejaculation seems particularly amenable to a self-directed format, since it enjoys a high probability of success, is straightforward, and can be easily standardized.

If empirically validated, this type of program has the potential to make effective, yet inexpensive, treatment available to the general public. Unfortunately, data have been reported for only two of the sexual self-help programs (Kass & Strauss, 1975; Lowe & Mikulas, 1975), and only one has been submitted to any controlled evaluation. Lowe and Mikulas (1975) found that couples using a written program in conjunction with telephone contact (of unspecified length) with a therapist reported significant improvement in latency to ejaculation, whereas untreated control couples reported no change in latency. Follow-up data were not reported.

Zeiss (1977) reported on two couples who successfully used a second self-directed program (Zeiss & Zeiss, 1978) in conjunction with minimal phone contact with a therapist (less than 1 hour total per couple) to treat premature ejaculation difficulties. An 8-month follow-up revealed that therapeutic gains maintained.

Even though these studies suggest that self-directed treatment for premature ejaculation may be effective, there were sufficient problems with each to preclude firm conclusions. Lowe and Mikulas (1975) relied only on subjective estimates of ejaculatory latency and did not report on the maintenance of treatment gains or on the amount of telephone contact required. Zeiss (1977) included objective indices of ejaculatory latency, provided an 8-month follow-up, and reported the amount of therapist contact involved, but firm conclusions cannot be drawn from uncontrolled case reports.

Finally, it is unclear exactly what successful treatment of premature ejaculation does. Both the Masters and Johnson (1970) intervention and the Semans (1956) intervention involve teaching clients to interrupt the arousal process, either by applying the

squeeze or by ceasing all stimulation of the penis. It is not clear, however, if greater latency to ejaculation results simply from repeated interruption of the arousal process with squeezes or cessation of stimulation or if practice with these exercises enables clients to tolerate greater and lengthier stimulation without the need for interruption of arousal.

To address these issues, three treatment conditions were compared. One group was treated in a "standard" sex therapy format in which both partners of the dysfunctional couple regularly saw a therapist in the clinic. Two treatment conditions used the Zeiss and Zeiss (1978) manual. Treatment was entirely self-administered for one of these groups; the other group had minimal therapist contact (phone) to increase the probability of successful outcome.

Method

Clients

Twenty heterosexual couples with self-defined premature ejaculation difficulties began treatment. These client couples also met the following criteria: (a) mean timed ejaculatory latency was less than 5 minutes, (b) both partners agreed to participate in the treatment program, (c) the female had no severe gynecological problems, and (d) the couple had experienced the problem for at least 6 months. The 20 couples who began treatment were randomly assigned to one of the three treatment conditions.

Of the 20 couples who began treatment, 2 (1 in each self-directed treatment condition) completed treatment and verbally reported success but failed to complete posttreatment assessment. Because there were no posttreatment data for these couples, they were excluded from all data analyses and further consideration. Data are reported on 18 client couples, 6 in each treatment condition. Demographic and descriptive data for the three conditions are presented in Table 1. One-way analyses of variance revealed that couples in the three treatment conditions did not differ significantly on any of these variables before treatment.

Therapists

Therapists included three male graduate students in counseling psychology and two female BA-level paraprofessionals, in addition to the author (a male graduate student in clinical psychology). Only the author had previously treated couples with sexual dysfunction; the other therapists were trained and supervised by him.

Table 1

Descriptive Statistical Means for the Three Treatment Groups

Variable	Standard treatment	Phone contact	No contact
Age ^a			
Males	28.5	33.2	30.8
Females	28.8	27.8	29.8
Education ^a			
Males	15.3	14.3	14.2
Females	14.8	14.5	14.5
Length of marriage or cohabitation ^a	3.2	5.5	6.8
No. children	1.8	2.1	1.2
Gross yearly income	\$8,883.33	\$12,900.00	\$14,216.67
Duration of premature ejaculation problem ^a	7.7	14.7	9.7
Mean timed ejaculatory latency (pretreatment) ^b	83.0	118.8	54.2
Marital Adjustment Test (Locke & Wallace, 1959)			
Males	104.2	114.5	110.2
Females	110.7	105.3	100.2

^a In years.^b In seconds.

Procedure

During pretreatment assessment, all couples completed the Locke and Wallace (1959) Marital Adjustment Test and a sexual background inventory. After receiving general information about the study and deciding to participate, couples were asked to report the male's latency from intromission to ejaculation, timed by stopwatch, on two separate occasions of self-defined "normal intercourse."

At a second meeting with their therapist, couples paid a \$10 treatment fee and a refundable contingency deposit of \$30. Refund of the deposit was contingent on regular completion of assignments and completion of posttreatment assessment.

At the conclusion of treatment or after 15-20 weeks from the start of treatment (whichever was earlier), couples again completed the Marital Adjustment Test and the sexual background inventory. In addition, couples timed the male's ejaculatory latency (a) twice as they "normally" had intercourse, reporting the number of squeezes or pauses used each time, and (b) twice during intercourse without using either the squeeze or the pause.

Three months after the completion of posttreatment assessment, the Marital Adjustment Test and the sexual background inventory were mailed to the couples, and couples were asked to complete them.

Treatments

Self-administered manual (no contact). Client couples in this condition were given the treatment

manual (Zeiss & Zeiss, 1978)¹ at their second meeting with their therapist. This manual describes a 12-week training program incorporating the Masters and Johnson (1970) squeeze technique and the Semans (1956) pause technique. During each week, about 3 hours of specific sexual and talking activities are assigned. The talking assignments are intended to facilitate a couple's communication about sexuality and to enhance verbal intimacy. Each week's lesson is followed by a trouble-shooting guide that discusses problems commonly encountered during that week's activities. After an introductory discussion of premature ejaculation and the use of the treatment program, the manual instructs couples in the squeeze and the pause procedures, as well as in sensate focus (Masters & Johnson, 1970), or "pleasuring," exercises. Subsequently, the squeeze and pause are incorporated into the pleasuring exercises to delay ejaculation and to prolong sexual activity in a graduated sequence of sexual interactions. The pleasuring exercises and extravaginal stimulation systematically approach "normal" intercourse, with the expectation that a couple will learn to enjoy ejaculatory control during sexual activity of their own choosing before completion of the program.

Couples in the no-contact condition were encour-

¹ This is a revised version of the manual actually used in the study. The version used by clients in the study was a prepublication draft. This preliminary version can be made available, at cost, to those requesting it from the author.

aged to work diligently and regularly at the treatment program and were told to contact the clinic on completion of the program or earlier if they ran into insurmountable problems. (None actually requested help during treatment.) After 15–20 weeks, all couples were asked to complete the posttreatment assessment.

Self-administered manual with minimal therapist contact (phone contact). This condition involved brief weekly phone contact with the therapist at prearranged times. Total phone contact averaged 71.5 minutes per couple, or about 6 minutes each week. Phone contact was supportive in nature and served primarily to check on couples' progress, to congratulate their successes, to provide encouragement when couples felt discouraged, and to help them resolve minor problems when they arose.

Therapist-administered treatment (standard treatment). In this condition, client couples were seen once weekly, at the University Psychology Clinic, by a therapist. The same specific treatment procedures, exercises, and schedules were used as are described in the manual; treatment was limited to 12–20 therapy sessions of no more than 1 hour each for no more than 20 weeks. The manual was not given to these clients.

Results

Posttreatment

The 18 couples who began treatment and provided posttreatment data were classified as successful or unsuccessful in treatment according to the following criteria: (a) Mean timed latency in normal intercourse must be greater than 5 minutes or must have improved by 3 or more minutes from pretest, and (b) both partners must report improvement in satisfaction with ejaculatory latency. By these criteria, 11 of the 18 couples were successful in treatment. All couples in standard treatment and 5 of 6 couples in the minimal phone contact condition successfully overcame their premature ejaculation problems, whereas none of the 6 couples using the manual entirely on their own was successful. Most couples dropped out early in treatment, typically stating that they were too busy with other activities to devote enough time to the treatment and that they would return to it when their busy lives allowed. Couples who dropped out of treatment were offered treatment in one of the successful treatment conditions, but none accepted the offer. Only 1 couple in the no-contact condition completed treatment. This couple was rated as unsuccessful, despite improvement in latency, because the

Table 2
Mean Timed Ejaculatory Latencies for Successfully Treated Couples

Treatment	Pre-treatment	Post-treatment
Standard ^a	1 min 23 sec	10 min 48 sec
Phone contact ^b	1 min 54 sec	10 min

Note. Timed latency data were not collected at follow-up.

^a $n = 6$.

^b $n = 5$.

female partner did not report improvement in her satisfaction with the male's ejaculatory latency.

Additional analyses compared the effectiveness of standard in-clinic treatment and self-directed treatment with phone contact, using two-factor repeated measures analyses of variance (unweighted-means solution) comparing pretreatment and posttreatment information. The first of these analyses examined mean timed ejaculatory latencies. This analysis showed a strong main effect from pretreatment to posttreatment, $F(1, 9) = 21.24$, $p < .005$ (means appear in Table 2). There were no significant differences between treatments or interaction effects. A similar analysis of variance was computed using couples' estimates of ejaculatory latency (averaged across the two partners). This analysis also indicates a significant main pre-post effect, $F(1, 9) = 66.74$, $p < .001$, and no interaction or differences between treatments. Mean latency estimates are shown in Table 3.

Analyses of variance revealed no significant change on the Marital Adjustment Test (Locke & Wallace, 1959) either for males or for females. The effects of successful treatment for premature ejaculation on the general sexual relationship were assessed through consideration of the sex quality composite scores. This index is composed of questions from the Sexual Background Inventory (pre or post), chosen on an intuitive, *a priori* basis.² Analyses of variance revealed signifi-

² For males, the index consists of items concerning frequency of intercourse, usual length of foreplay, frequency of premature ejaculation, satisfaction with ejaculatory latency, anxiety over ejacu-

Table 3
Mean Estimates of Ejaculatory Latency for Successfully Treated Couples

Treatment	Pre	Post	Follow-up
Standard	1 min 23 sec ^a	8 min 22 sec ^a	4 min 52 sec ^b
Phone contact	2 min 8 sec ^b	8 min 18 sec ^b	3 min 55 sec ^a

Note. Because of the small n , follow-up data were not included in analyses of variance.

^a $n = 6$.

^b $n = 5$.

^c $n = 3$.

cant main pre-post effects on this variable, both for males, $F(1,9) = 91.71$, $p < .001$, and for females, $F(1,9) = 41.59$, $p < .001$. There were no significant differences between treatments or interactions for either sex. Mean scores on this variable are shown in Table 4.

Maintenance

Of the 11 couples who were successfully treated, 8 returned follow-up data at a mean of 4.4 months (range = 3-9 months) after posttreatment assessment. Of the 8 couples who provided follow-up data, 4 were classified as successfully treated at the time of follow-up on the basis of the two criteria described previously. Three of these couples had received standard in-clinic treatment for premature ejaculation, and 1 had self-directed treatment with phone contact. The other 4 couples (2 in each condition) were classed as no longer successful either because estimates of ejaculatory latency (averaged across the two partners) no longer reached criterion and/or because one of the partners reported

no improvement (over pretest) in their satisfaction with ejaculatory latency.

Of all couples completing the follow-up assessment, two couples reported the same latency as at posttreatment, and six couples estimated their latency to be less than at posttreatment. The differences between the two mean estimates (8 min 11 sec at posttreatment and 4 min 31 sec at follow-up) was significant by a t test for paired observations, $t(7) = 3.81$, $p < .005$, indicating that ejaculatory latencies did decrease in the months following treatment. However, despite the decrease in latencies, follow-up estimates of latency remained significantly greater than pretreatment estimates ($M = 1$ min 43 sec at pretreatment), $t(7) = 3.47$, $p < .01$. A correlation between the two measures (.58) indicated that posttreatment latencies were largely predictive of follow-up latencies.

The sex quality composite scores for these

Table 4
Mean Sex Quality Composite Scores for Successfully Treated Couples

Treatment	Pre	Post	Follow-up
Males			
Standard	23.8 ^a	42.4 ^a	40.8 ^b
Phone contact	25.2 ^a	38.5 ^a	30.1 ^d
Females			
Standard	25.8 ^a	43.1 ^a	37.8 ^a
Phone contact	28.7 ^a	47.7 ^a	34.7 ^d

Note. Because of the small n , follow-up data were not included in analyses of variance.

^a $n = 6$.

^b $n = 4$.

^c $n = 5$.

^d $n = 3$.

latory control, satisfaction with the sexual relationship, and response to sexual advances by mate. The index for females consisted of the same seven items plus an item regarding the frequency of orgasm in intercourse. Responses to each question were assigned numerical equivalents on a 9-point scale (0-8), with equal intervals between the two extremes. Larger numbers reflected higher quality of the sexual relationship. Thus, the maximum possible score for males on this index was 56, for females, it was 64. Item intercorrelations for the sex quality composite at pretest are available on request from the author (only 3 of 21 item intercorrelations for males and only 3 of 28 for females are significant). Thus, the items contributing to the composite mean are largely independent.

eight couples were also examined. The pattern was the same as with ejaculatory latency—improvement with treatment, followed by a moderate deterioration by the time of follow-up, both for males and for females. Means for males were 26.36 at pretest, 42.57 at posttest, and 36.18 at follow-up; means for females were 28.50 at pretest, 45.50 at posttest, and 36.75 at follow-up. t tests for paired observations indicated that sex quality scores dropped significantly between posttreatment and follow-up assessments for both sexes: For males, $t(6) = 2.33$, $p < .05$; for females, $t(7) = 3.06$, $p < .01$. Despite this decrease, the quality of sex at follow-up remained significantly greater than the quality of sex before treatment for both sexes: For males, $t(6) = 2.80$, $p < .025$; for females, $t(7) = 3.53$, $p < .005$. It was also found that the posttreatment sex quality composite for males was highly predictive of follow-up ejaculatory latency estimates, $r(5) = .87$, $p < .01$, whereas the composite for females was not ($r = -.10$).

Mechanism of Successful Treatment

To examine the mechanism by which treatment for premature ejaculation works, couples at posttreatment were asked to time their ejaculatory latencies first as intercourse normally occurred and then without using any squeezes or pauses; 10 couples provided these data. Mean latency was 10 min 33 sec as intercourse normally occurred and 4 min 21 sec with no squeezes or pauses, $t(9) = 2.90$, $p < .01$. Mean posttreatment latency without squeezes or pauses was significantly greater than mean pretreatment latency, which was 1 min 30 sec for these 10 couples, $t(9) = 3.82$, $p < .005$. Thus, although treatment was most effective when couples continued to use the squeeze or pause, there was also a significant effect of treatment on simple ejaculatory latency when squeezes and pauses were not used.

Discussion

The results of this study indicate, in a population with fairly stable relationships and long histories of premature ejaculation, that couples can successfully treat themselves

using a written instructional guide *if they maintain minimal contact with a therapist*. In this instance, minimal therapist contact was by telephone, averaged 6 minutes per week, and generally was not specifically therapeutic in nature. Rather, phone contacts served primarily to check on, encourage, and congratulate treatment progress. Without the non-specific therapeutic effects of ongoing contact with a therapist, couples failed to complete treatment successfully, generally because of ceasing to follow the treatment plan. These results were consistent across a variety of measures, including timed ejaculatory latency, estimated latency, and the sex quality composite. There did not seem to be indications of concomitant improvement in the general marital relationship.

The brevity and nature of phone contacts with minimal therapist contact couples suggest that the treatment exercises outlined by the Zeiss and Zeiss (1978) manual are sufficient for successful treatment when followed. However, evidence from the no-contact condition indicates that an external source of encouragement and motivation is probably necessary in order for couples to follow through with treatment exercises. Graduate student and relatively inexperienced paraprofessional therapists filled the therapist role in this study; the same function could probably be fulfilled by others with minimal training in the treatment of sexual dysfunction. These might include physicians, psychologists, mental health paraprofessionals, or the clergy.

The success of treatment with minimal therapist contact suggests that the cost of a standard, in-clinic therapist-administered treatment for premature ejaculation may no longer be justifiable for those couples with well-defined and simple premature ejaculation difficulties. Rather, couples could be offered the use of the self-directed treatment program in conjunction with minimal contact with a therapist. In this framework, intensive professional counseling could be reserved for those couples who fail to succeed with the minimal contact approach or for couples whose premature ejaculation difficulties are compounded by other clinical problems.

A recent review of the behavioral self-help

literature (Glasgow & Rosen, 1978) indicates that dropout problems similar to those found in this study have been noted with self-directed treatments in other areas such as fear reduction (Clark, 1973; Marshall, Presse, & Andrews, 1976; Phillips, Johnson, & Geyer, 1972; Rosen, Glasgow, & Barrera, 1976), weight reduction (Hanson, Borden, Hall, & Hall, 1976; Mahoney, Moura, & Wade, 1973), and study skills training (Beneke & Harris, 1972; Harris & Ream, 1972). Many of these studies noted dropout problems with minimal therapist contact applications, as well as with strictly self-administered treatment applications. In the current study, the dropout problem was restricted to the self-administered condition but persisted even with the existence of monetary contingency deposits.

Previous work on the treatment of premature ejaculation has not specified the mechanism by which treatment is effective, although Levine (1976) has reported that unlimited control is often not developed. An analysis of posttreatment data indicated that greatest improvement in ejaculatory control occurred when couples continued to use the squeeze or pause to delay ejaculation, but significant improvement in latency to ejaculation also occurred when couples used neither technique to lengthen intercourse. This finding concurs with Levine's conclusions.

By the criteria used to define success, half of the successfully treated couples for whom there were follow-up data could no longer be classified as successfully treated at follow-up. The reasons for this decrease in latency and general quality of sex are not clear, although it may be that there is a glow of success that tends to deteriorate in the months following treatment for sexual dysfunction, or couples may have simply decreased their use of the squeeze or pause after completing treatment. It should be noted that although estimated latencies and the sex quality composite both decreased following completion of treatment, couples remained significantly improved on these variables at follow-up, compared to pretreatment assessment.

Also of interest is the finding that post-treatment ejaculatory latency is predictive of ejaculatory latency at follow-up. This suggests that even though there was a general

decrease in latencies from posttreatment to follow-up, the decrease occurred across most couples, resulting in similar ranking of improvement at follow-up. Even more powerfully predictive of ejaculatory latency at follow-up was the posttreatment sex quality composite for men. The composite for women was not predictive of follow-up results. This indicates that improvement in latency tends to maintain when the male partner is satisfied with the sexual relationship; if he is less satisfied with the sexual relationship, improvements in his ejaculatory latency fail to maintain.

This finding may shed some light on a factor that has perplexed sex therapists at this clinic for years. Many couples who have been successfully treated for sexual dysfunction have improved on their target problem during treatment, but after the termination of treatment, the frequency of sexual activity tends to revert to very low pretreatment levels, in spite of reports that improvement in the target areas has maintained. That is, the man retains his newly learned ejaculatory control or the woman retains her capacity for orgasm (LoPiccolo, Note 1). This contradiction is perplexing because of the presumed reinforcement value of sexual activity in general and orgasm in particular.

Although couples are referred with specific dysfunctions, and although we as behavior therapists eagerly label and treat their dysfunctions, it may be that resolution of the specific dysfunctions does not adequately resolve these couples' unspecified needs for a general improvement of the sexual relationship. Thus, in addition to resolving specific dysfunctions, therapy should probably concern itself with all aspects of the sexual relationship. If the general relationship does not improve along with the target behavior, improvements in that target behavior will not necessarily maintain.

Reference Note

1. LoPiccolo, J. Personal communication, August 2, 1977.

References

- Aycock, L. The medical management of premature ejaculation. *Journal of Urology*, 1949, 62, 361-362.

- Barbach, L. G. *For yourself: The fulfillment of female sexuality*. New York: Doubleday, 1975.
- Beneke, W. M., & Harris, M. B. Teaching self-control of study behavior. *Behaviour Research and Therapy*, 1972, 10, 35-41.
- Bennett, D. Treatment of ejaculatio praecox with monoamine-oxidase inhibitors. *Lancet*, 1961, 2, 1309.
- Boneff, A. N. Topical treatment of chronic prostatitis and premature ejaculation. *International Urology and Nephrology*, 1971, 4, 183-186.
- Clark, F. Self-administered desensitization. *Behaviour Research and Therapy*, 1973, 11, 335-338.
- Clarke, M., & Parry, L. Premature ejaculation treated by the dual sex team method of Masters and Johnson. *Australian and New Zealand Journal of Psychiatry*, 1973, 7, 200-205.
- Cooper, A. J. Clinical and therapeutic studies in premature ejaculation. *Comprehensive Psychiatry*, 1969, 10, 285-295.
- Friedman, D. The treatment of impotence by brieft relaxation therapy. *Behaviour Research and Therapy*, 1968, 6, 257-261.
- Glasgow, R. E., & Rosen, G. M. Behavioral bibliography: A review of self-help behavior therapy manuals. *Psychological Bulletin*, 1978, 85, 1-24.
- Hanson, R. W., Borden, B. L., Hall, S. M., & Hall, R. G. Use of programmed instruction in teaching self-management skills to overweight adults. *Behaviour Therapy*, 1976, 7, 366-373.
- Harris, M. B., & Ream, F. A program to improve study habits of high school students. *Psychology in the Schools*, 1972, 9, 325-330.
- Heiman, J., LoPiccolo, L., & LoPiccolo, J. *Becoming orgasmic: A sexual growth program for women*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.
- Kaplan, H. S. *The new sex therapy*. New York: Brunner/Mazel, 1974.
- Kaplan, H. S., Kohl, R. N., Pomeroy, W. B., Offit, A. K., & Hogan, B. Group treatment of premature ejaculation. *Archives of Sexual Behavior*, 1974, 3, 443-452.
- Kass, D. J., & Strauss, F. *Sex therapy at home*. New York: Simon & Schuster, 1975.
- Kline-Graber, G., & Graber, B. *Woman's orgasm: A guide to sexual satisfaction*. New York: Bobbs-Merrill, 1975.
- Levine, S. B. Marital sexual dysfunction: Ejaculation disturbances. *Annals of Internal Medicine*, 1976, 84, 575-579.
- Locke, H. J., & Wallace, K. M. Short marital-adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living*, 1959, 21, 251-255.
- Lowe, J. C., & Mikulas, W. L. Use of written material in learning self-control of premature ejaculation. *Psychological Reports*, 1975, 37, 295-298.
- Mahoney, M. J., Moura, N. G. M., & Wade, T. C. Relative efficacy of self-reward, self-punishment, and self-monitoring techniques for weight loss. *Journal of Consulting and Clinical Psychology*, 1973, 40, 404-407.
- Marshall, W. L., Presse, L., & Andrews, W. R. A self-administered program for public speaking anxiety. *Behaviour Research and Therapy*, 1976, 14, 33-40.
- Masters, W. H., & Johnson, V. E. *Human sexual inadequacy*. Boston: Little, Brown, 1970.
- Mellgren, A. Treatment of ejaculatio praecox with thioridazine. *Psychotherapy and Psychosomatics*, 1967, 15, 454-460.
- Phillips, R. E., Johnson, G. D., & Geyer, A. Self-administered systematic desensitization. *Behaviour Research and Therapy*, 1972, 10, 93-96.
- Rosen, G. M., Glasgow, R. E., & Barrera, M., Jr. A controlled study to assess the clinical efficacy of totally self-administered systematic desensitization. *Journal of Consulting and Clinical Psychology*, 1976, 44, 208-217.
- Schapiro, B. Premature ejaculation: A review of 1,130 cases. *Journal of Urology*, 1943, 50, 374-379.
- Semans, J. H. Premature ejaculation: A new approach. *Southern Medical Journal*, 1956, 49, 353-358.
- Vandervoort, H. E., & McIlvenna, T. *You can last longer*. San Francisco: Multi Media Resource Center, 1972.
- Wolpe, J. Reciprocal inhibition as the main basis of psychotherapeutic effects. *Archives of Neurology and Psychiatry*, 1954, 72, 205-226.
- Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford, Calif.: Stanford University Press, 1958.
- Zeiss, R. A. Self-directed treatment for premature ejaculation: Preliminary case reports. *Journal of Behavior Therapy and Experimental Psychiatry*, 1977, 8, 87-91.
- Zeiss, R. A., Christensen, A., & Levine, A. G. Treatment for premature ejaculation through male only groups. *Journal of Sex and Marital Therapy*, 1978, 4, 139-143.
- Zeiss, R. A., & Zeiss, A. M. *Prolong your pleasure*. New York: Pocket Books, 1978.
- Zilbergeld, B. Group treatment of sexual dysfunction in men without partners. *Journal of Sex and Marital Therapy*, 1975, 1, 204-214.

Therapeutic Effectiveness of Setting and Monitoring Goals

Russell R. Hart

Copper Mountain Mental Health Center, Salt Lake City, Utah

This study investigated the therapeutic effects of goal-setting strategies on patients whose behaviors were disintegrating but not fragmented or disoriented. Two techniques of goal setting were introduced into the therapeutic interaction to effect greater beneficial changes in patient attainment of goals: (a) collaboration of patients, therapists, and "collaterals" (other persons significant to the patient, such as spouse or probation officer) on 3-month treatment goals and (b) weekly monitoring with structured feedback. Sixteen outpatients were randomly assigned to a treatment group for which the Behavioral Monitoring Progress Record was used. The Behavioral Monitoring Progress Record is a method in which the patient and therapist collaborate in setting and monitoring weekly therapeutic goals. Sixteen other outpatients were randomly assigned to the same individual therapy but without weekly goal setting or monitoring. Collateral persons and behavioral criteria were used at intake for collecting information in each problem area and were used at follow-up to validate the patient's self-report. A score was used as an index of change. All patients improved from the time of intake to the time of the follow-up interview, eight individual therapy sessions later. There was greater success in the attainment of goals for patients using the Behavioral Monitoring Progress Record than for patients in the group not using this format.

Because of the increasing emphasis on accountability within community mental health centers, Kiresuk (1973) has proposed a method for evaluating treatment effectiveness in psychotherapy using goal attainment scaling procedures. Goal attainment scaling is a process in which therapeutic goals are set for, by, or with the subject. The possible levels of predicted attainment of these goals are scaled from the least favorable to the most favorable outcomes in behavioral terms.

Although the purpose of Kiresuk's work was to develop a method of quantifying behavioral change, the process of setting and scaling goals appeared to have potential as a therapeutic procedure.

The major objective in this study was to evaluate the therapeutic effectiveness of setting goals in behavioral terms while monitoring the subject's progress in attaining these

goals. To accomplish this objective, the answer to two questions were sought: Did behavior change occur from intake to follow-up evaluation? and Was there a difference in behavior change between the monitored and control groups?

Method

The subjects were 32 adult patients at a community mental health center whose mental functioning was not sufficiently impaired to require hospitalization. The decision to include a patient depended on the answers to two essential questions: (a) Is short-term (3 months) individual psychotherapy clinically appropriate treatment for the patient? and (b) Is the patient capable of participating meaningfully and responsibly in a therapy program?

Instruments

The instruments used in this study were the Goal Attainment Scale (GAS; Kiresuk & Sherman, 1968), the Behavioral Monitoring Progress Record, and the Follow-up Interview Schedule.

Requests for reprints should be sent to Russell R. Hart, 7340 Comstock Circle, Salt Lake City, Utah 84121.

BEHAVIORAL MONITORING PROGRESS RECORD

NAME John Doe THERAPIST Smith DATE OF 1ST SESSION 3/1/76

MAJOR PROBLEMS &/OR COMPLAINTS				
SEE SCALE HEADINGS ON FOLLOW-UP GUIDE	UNHAPPY WITH PRESENT EMPLOYER	FEELS DISHONEST, PHONY, PLAYING GAMES CAN'T BE SELF	DON'T STAND UP FOR RIGHTS	AVIODS INTERPERSONAL RELATIONSHIPS

4-WEEK GOALS

PROJECT FROM CLIENT STATUS AT INTAKE	FIND NEW EMPLOYMENT THAT IS STIMULATING AND HAS CAREER OPPORTUNITIES	TO BEGIN TO FEEL IT'S SAFE TO SAY WHAT YOU FEEL WITH SOME PEOPLE	BE ASSERTIVE AT LEAST ONCE A WEEK	TO BE INVOLVED IN ONE OR TWO FRIENDSHIPS OR CONFIDENCE
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

BE SPECIFIC, OBSERVABLE &/OR TASK-ORIENTED

WEEKLY GOALS

WEEK OF	GOAL:	INVESTIGATE JOB MARKET	EXPRESS CONCERN LIVING AT HOME	TO ASSERT SELF	SOCIALIZE WITH GIRLS AT CHURCH
	METHOD:	TWO JOB INTERVIEWS	COMMUNICATION - TIME AND PLACE	RETURN WRONG SIZE DRESS	GO ON DATING
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WEEK OF	GOAL:	INVESTIGATE JOB MARKET	DISCUSS AND IDENTIFY TWO PROBLEMS LIVING AT HOME	TO ASSERT SELF	BECOME ACQUAINTED WITH MARGE
	METHOD:	TWO OR MORE JOB INTERVIEWS	COMMUNICATION - BRING FATHER IN	COLLECT \$30 LOAN FROM "MERYL"	CALL MARGE FOR LUNCH DATE
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WEEK OF	GOAL:	REVIEW LIST OF QUESTIONS WITH MRS. J	EXPRESS ANGER OPENLY - SAY WHAT YOU THINK & FEEL	TO ASSERT SELF	SAY WHAT YOU THINK & FEEL
	METHOD:	SET UP APPOINTMENT	"LET IT OUT"	REQUEST \$600/MO. IN SALARY FROM MRS. J.	COMMUNICATION
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
WEEK OF	GOAL:	CHOOSE BETWEEN TWO ATTRACTIVE JOBS	SAME AS ABOVE	TO ASSERT SELF	GO OUT ON DOUBLE DATE WITH MARGE & PERRY
	METHOD:	THERAPY - RANK, PRIORITIZE & NET. MAKE DECISION	SAME AS ABOVE	LD. 3 MAJOR INCIDENTS LAST WK. IN WHICH YOU ASSERTED SELF	CALL MARGE FOR OR ON DOUBLE DATE
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

NOTES:

Figure 1. Behavioral Monitoring Progress Record.

The Behavioral Monitoring Progress Record (BMPR). The BMPR was designed to increase the therapeutic movement of the patient by having patients set goals and report on their progress at each therapy session (see Figure 1). A predicted 4-week goal was set with the patient, and successive weekly goal approximations were also determined. Within each problem area, a weekly goal and method of attainment was specified. These methods were carefully defined behavioral "assignments." The patient and therapist jointly assessed the degree of attainment of each goal. For example, unassertiveness might be a problem area. The 4-week goal might be to be assertive on seven occasions during the week; the method for the first weekly goal might be to stand up for one's rights once in a restaurant, once with the boss, and once with one's spouse. Hence, a most important emphasis was on setting weekly goals that were observable, definable, and measurable and methods that were structured in a step-by-step, reasonable,

and realistic manner and collaborated and monitored weekly by both patient and therapist.

The Follow-up Interview Schedule. The GAS Follow-up Interview Schedule (adapted from Garwick, 1974a), is a structured interview that allows for quantitative assessment of patient functioning within the problem areas for which goals have been set.

Patients who were selected as subjects in the study were asked to complete the "guide to goals" (Garwick, 1974b). Then an "intake history" was prepared on the basis of two interview sessions.

At a third session, the patient and an experienced clinician "scaler," trained in goal attainment scaling, collaboratively prepared a follow-up guide. The scaler interviewed "collateral" persons significant to the patient's problem areas. The scaler initially interviewed the collateral within each problem area that the patient presented and in a separate interview with the patient collaboratively constructed the goal attainment follow-up guide.

The construction of the follow-up guide consisted of setting treatment goals and predicting five levels of goal attainment with an "expected" level of attainment by the eighth therapy session. The patients were then randomly assigned to the two treatment groups and to the psychotherapists for individual counseling.

The monitored group received individual therapy and weekly goals using a structured feedback technique (BMPR). The nonmonitored group received individual therapy *without* setting goals.

The Follow-up Interview Schedule ratings were completed after eight therapy sessions within a 3-month period by one of four master's level psychiatric nurses. Separate posttest scores were recorded for both patients and collaterals.

The collateral person was a source of external validation of the patient's self-report. Validation included identification and definition of the patient's problems at intake (pretest score) and input as to the level of functioning on the attainment level of the follow-up guide at the follow-up evaluation (collateral posttest score).

Results

The purpose of this study was to determine what changes, if any, occurred from the time of intake to the time of follow-up between two treatment groups, only one of which used a structured feedback technique for collaboration and monitoring goals.

The following hypotheses were used to examine the therapeutic effectiveness of goal attainment scaling:

Hypothesis 1. There is no difference in the mean Goal Attainment scores for the pretest, the posttest, and the collateral posttest.

Hypothesis 2. There is no difference in the means on the goal attainment scores of the treatment and control groups.

Hypothesis 3. There is no interaction between testings and treatment groups on the Goal Attainment scores.

To test these hypotheses, a two-way analysis of variance (see Table 1) with repeated measures on one factor (Winer, 1962) was used.

Mean Total Pretest Versus Patient Posttest Versus Collateral Posttest Goal Attainment Scores

The mean totals of the patient (49.16) and collateral (50.15) after therapy were more than twice as great as the mean total pretherapy score (23.29). As shown in Table 1, a significant treatment effect was found for pretherapy versus patient posttherapy versus collateral posttherapy Goal Attainment scores ($F = 327.66, p = .01$).

Mean Total Control Group Versus Monitored Group Attainment Scores

The mean total Goal Attainment score for the monitored group (45.73) was much greater than the mean total Goal Attainment score for the control group (36.01). The monitored patients attained significantly higher scores than the control patients ($F = 32.43, p = .01$).

Interaction

Control Group Versus Monitored Group

Both the patient (56.82) and collateral (57.72) posttherapy scores of the monitored

Table 1

Summary Table for Two-Way Analysis of Variance for Control and Monitored Groups

Source of variation	SS	df	MS	F
Between subjects	4,370.05	31		
Monitored vs. control (A)	2,270.20	1	2,270.20	32.43*
Subjects within groups	2,099.84	30	70.00	
Within subjects	17,656.01	64		
Pre, post, and collateral observations (B)	14,847.10	2	7,423.55	327.66*
A \times B	1,449.53	2	724.80	31.99*
B \times subjects within groups	1,359.38	60	22.70	

Note. $N = 32$.

* $p < .01$.

Table 2
Mean Initial and Follow-Up Goal Attainment Scores

Item	<i>M</i> initial scores (all scales)	<i>M</i> score based on patient report at 3 months	<i>M</i> score based on ratings by collaterals at 3 months	Total (between groups)
Initial follow-up guide only (control group)	23.92	41.51	42.59	36.01
Weekly goals review plus initial follow-up guide (monitored group)	22.66	56.82	57.72	45.73
Total (within groups)	23.29	49.16	50.15	40.87

group were significantly higher than both the patient (41.51) and collateral (42.59) posttherapy scores of the control group ($F = 31.99, p = .01$).

Table 2 represents the mean Goal Attainment scores at pretherapy, patient posttherapy, and collateral posttherapy interactions. There were differences between testings and treatment groups on the Goal Attainment scores. Patients in the monitored group had significantly higher mean scores in both posttestings on patient reports and collateral ratings than did patients in the control group, whereas mean initial scores for both treatment groups were similar.

Discussion

The significant change in the mean Goal Attainment scores from pretherapy to posttherapy suggests that positive therapeutic changes occurred during treatment. Undesirable behaviors decreased, and desirable behaviors increased. The treatment goals that were selected and the levels of outcome that were achieved demonstrate that reasonable and realistic goals were set and accomplished with these patients.

The significant difference between the mean Goal Attainment scores for patients in the monitored group and the control group suggests that greater beneficial changes in patient attainment of goals were effected in the monitored group using a structured format of patient-therapist collaboration on weekly goals than when the guide to goals was used without weekly goal setting.

It appears that goal-setting strategies, when used in a procedural format, allow the

patient and therapist to be more responsible and accountable in assuring the process of appraisal and improvement of the quality. Behavior changed gradually and improved when the patient and therapist had the feedback of the established goals of the prior week and methods to accomplish these goals. Feedback provided both the patient and therapist with information on treatment decision and outcome and provided alternative approaches for future goal attainment strategy.

These results indicate that the goal attainment model with periodic monitoring may be useful in the therapeutic process to collect information, as an aid in organizing and recording the process of therapy, to design treatment for outpatients, as an outcome effectiveness measure, to evaluate therapeutic progress, and to provide new data for setting additional therapeutic goals.

References

- Garwick, G. *An introduction to reliability and goal attainment scaling methodology* (Program evaluation project report). Minneapolis, Minn.: Program Evaluation Resource Center, 1974 (a).
- Garwick, G. Recent findings on the use of goal-setting in human services. *Goal Attainment Review*, 1974, 1, 1-4. (b)
- Kiresuk, T. J. Goal attainment scaling at a county mental health service. *Evaluation*, 1973, 1, 12-18.
- Kiresuk, T. J., & Sherman, R. Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal*, 1968, 4, 443-453.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Received July 18, 1977 ■

An Observational Approach to the Assessment of Anxiety in Young Children

Blair Glennon
Harvard University

John R. Weisz
University of North Carolina at Chapel Hill

Clinical research on anxiety has long relied on assessment techniques that may be inappropriate with young children (e.g., self-report inventories). The present article describes an alternative to such techniques—a scale using observational methodology. To assess the reliability and validity of this instrument, the Preschool Observational Scale of Anxiety (POSA), preschoolers were observed and scored on the scale during two test sessions. Session 1, with mothers absent, was expected to provoke relatively high anxiety; Session 2, with mothers present, was expected to provoke minimal anxiety. Total POSA scores assigned by two independent judges correlated .78 ($p < .001$), with highly significant interjudge correlations for most of the scale items. Regarding the validity of the instrument, it was found that (a) POSA scores were significantly correlated with teachers' and parents' inventory ratings of children's anxiety (all $ps < .01$), and (b) children's POSA scores were significantly higher in Session 1 than Session 2 ($p < .01$). The findings suggest that the POSA may provide a means of assessing situationally induced anxiety in children who are too young to accurately report their internal states.

Anxiety has long been a topic of central importance in clinical research and practice. Achenbach (1974) noted that among personality traits emerging from trait theories, "anxiety is perhaps the most frequently inferred and measured (p. 574)"; yet, demonstrably valid and reliable measurement of anxiety has been difficult to achieve, particularly among children. Especially acute is the need for accurate measures that can be used to assess specific situational effects on anxiety states in children (see Spielberger et al., 1972). Although many investigators recognize the need for such measures, particularly in research with children who are too young to

accurately report their own internal states, most researchers also recognize the difficulties that inhere in applying previous anxiety assessment techniques to children.

Historically, the principal methods for measuring anxiety have been physiological measurements, projective techniques, self-reports, and behavior ratings by observers. Through direct gauging of autonomic activity, physiological measures bypass the problems of subjective judgments. However, agreement among physiological measures is frequently poor, since people have different styles of autonomic responding (Phillips, Martin, & Meyers, 1972), and the physiological instruments may tap emotions such as anger or joy rather than anxiety (Lazarus, 1966). Furthermore, the unusual instruments used for physiological measures may distract subjects and make naturalistic observations difficult, particularly when the subjects are children.

Projective techniques have also been used to obtain ratings of generalized anxiety (McReynolds, 1968), but most projective techniques require individual administration as well as substantial time and expertise on the

The study was supported in part by a research grant from the College of Human Ecology, Cornell University.

The authors are grateful to Barbara Cesario, John Doris, Laura Harrison, Robert Loheyde, Linda Mack, and the staff, parents, and children of the Cornell Nursery School for their assistance in the conduct of this study.

Requests for reprints should be sent to John R. Weisz, Department of Psychology, University of North Carolina, Chapel Hill, North Carolina 27514.

part of the examiner, thus limiting their usefulness. Furthermore, McReynolds (1968) has reported that projective techniques "are generally found to be unrelated to inventory essays of anxiety" (p. 258); and Achenbach (1974), after reviewing the massive literature on projective techniques, has concluded that "there is little evidence for the reliability or validity of most of the interpretations made from them" (p. 604).

Three main types of self-reports have been used in assessing anxiety. First, direct self-ratings involve asking subjects specifically how anxious or nervous they feel. McReynolds (1968) has suggested that such self-ratings are most useful for measuring current levels of anxiety or changes in anxiety from one situation to another. A second type of self-report is the Adjective Check List, on which subjects indicate which of several adjectives (e.g., "jittery") best characterize their mood. Adjective Check Lists have been used to assess either current anxiety level or characteristic anxiety (McReynolds, 1968). The third type of self-report is the inventory method, with questionnaires designed to determine how subjects feel in a variety of situations. This method is usually aimed at assessing characteristic rather than current anxiety level. Inventories may focus on generalized anxiety (i.e., across a wide variety of situations) or anxiety related to particular types of situations (such as separations). Several investigators have noted drawbacks of the various self-report methods (see Spielberger, 1972); however, probably the most important drawback in the present context is that such methods assume both the ability and the willingness of subjects to correctly describe their own feelings. The first assumption is clearly tenuous with young children. And, with regard to the second assumption, there is evidence that many children give inaccurate reports about their anxiety due to defensiveness or social pressures (Sarason, Davidson, Lighthall, Waite, & Ruebush, 1960). Consequently, Sarason (1966) has stated that "the verbal response to our [children's self-report anxiety] scales may be telling us more about the self than the affect" (p. 79).

Behavior ratings of anxiety by observers may involve global, subjective judgments

about the subjects' apparent nervousness or ratings on many specific, concrete behaviors thought to indicate anxiety (e.g., stuttering). Such approaches appear to offer distinct advantages relative to those discussed above (see Spielberger et al., 1972). For instance, unlike physiological measures, observers' behavior ratings can be made unobtrusively; unlike projective techniques, they do not necessarily require trained clinicians; and unlike self-report measures, they do not rest on tenuous assumptions about the abilities and attitudes of the subjects. In addition, behavioral observations may be particularly useful with children, since they appear to disguise expression of their feelings less effectively than adults (Sarason, 1966).

The observers for the behavior rating scales of children's anxiety are sometimes teachers or parents, with the scales including items relevant to school or family situations (e.g., cries at bedtime). However, these particular scales entail a risk of distortion due to the parent's or teacher's lack of care or objectivity in observing, or to their bias or defensiveness about reporting on a child whom they know personally (Sarason et al., 1960). The observers for the behavior rating scales, however, may be trained observers (see Buss, Wiener, Durkee, & Baer, 1955), and the scales may comprise behaviors indicative of anxiety across situations, thus being more widely applicable and less subject to distortion by observers. The latter type of behavior rating scale would, in principle, permit detection of subtle relationships between specific events and ensuing anxiety in a way that satisfies methodological requirements and is developmentally appropriate as well (see Weisz, 1978).

It is true that different individuals may reveal their anxiety through different behaviors; also, different observers may have difficulty achieving high agreement in their behavior observations. But these problems of behavior ratings should be surmountable. Clearly defining the behaviors to be observed and increasing the training time of observers should reduce problems of low interrater agreement. And total frequency scores on rating scales including the gamut of behaviors suggestive of anxiety should give good indi-

cations of the relative anxiety of individuals with widely varying anxiety manifestations.

The most accurate behavior rating scale would require observers who are able to concentrate fully on the occurrence of the target behaviors and to record the occurrence of these behaviors for later frequency analysis. It should be noted that in the typical clinical interview situation, the clinician would not be able to meet these requirements. However, a second observer could independently record the indicators in clinical situations in which precise anxiety measurements are desired, or the clinicians themselves could use the indicators of behavior rating scales as an aid in making less exact, more global ratings of the client's anxiety. In the latter case, the clinicians would not be using the scale in the prescribed way but would probably add objectivity to their observations by reference to it.

In short, behavior observation measures of anxiety seem to have clear advantages over other types of anxiety measures and relatively minor drawbacks. But despite the potential usefulness of such measures, we have found little systematic research on behavior indicators of anxiety in children. In fact, there appears to be no carefully validated behavior rating scale for anxiety in children. Perhaps the closest approach is that of Grossman (1968), who used observer ratings of anxiety with 6-year-olds. However, his scale consisted of a mixture of specific objectively observable behaviors (e.g., nail-biting) and general indicators requiring subjective inference (e.g., "reactions suggesting that the child was frightened"); as described by the author, the scale appeared to contain only six items, thus showing limited sensitivity to the range of behaviors through which children may display anxiety. Validity data consisted of only two nonsignificant correlation coefficients relating the behavior scale to the General Anxiety Scale for Children (Sarason, Davidson, Lighthall, & Waite, 1958), and to our knowledge, no independent studies have been done to support the use of the scale as an anxiety measure. Melamed and Siegel (1975) used their Observer Rating Scale of Anxiety in a study on the reduction of anxiety through filmed modeling in 4- to 12-year-old

children facing surgery. Again, only very limited validity data on this scale are available. Also, only four items of the scale were published, and these represent an exceedingly modest sampling of the possible behavioral manifestations of anxiety.

In the present study, a much more detailed list of behavior items was assembled by means of a systematic search of the anxiety literature. This list was used to form the Preschool Observation Scale of Anxiety (POSA). The interobserver reliability of the POSA was assessed by two independent judges. The validity of the scale was determined (a) through assessment of its relation to three independent inventory measures of anxiety and (b) through an experimental manipulation of stressors. Significant positive relations between the POSA scores and the inventory scores were anticipated, but these relations were expected to be modest, since (a) the inventory measures and the POSA were designed to assess somewhat different aspects of anxiety, and (b) the inventory measures relied on parents' and teachers' judgments, whereas the POSA involved relatively objective observations by trained observers. The experimental manipulation involved a contrast between an initial mother-absent condition and a subsequent mother-present condition. Because the format and experimental procedure would be more familiar to the children in the second session, and because mothers would be present, it was predicted that children would show lower levels of anxiety in Session 2 than in Session 1, and thus score lower on the POSA.

Method

Subjects

The 36 Cornell University Nursery School children who formed the sample ranged in age from 32 to 59 months ($M = 47$ months; $SD = 7.7$ months). Mean Hollingshead (Note 1) socioeconomic status (SES) was 1.47 (1 = highest; 7 = lowest; $SD = .94$). There were 21 girls and 15 boys.

Procedure

The Preschool Observation Scale of Anxiety. The first step in devising the scale was a systematic search of the *Education Index* (1929-1976), *Psychological Abstracts* (1927-1976), *Resources in Education/Re-*

search in Education [1966 (year or origin) to 1976], and *Current Index to Journals in Education* [1969 (year of origin) to 1976], for studies using or mentioning behavioral indicators of anxiety in children or adults. Next, three child-clinical psychologists examined a list of items based on the literature search to suggest additions, to make minor modifications, and to give their final approval to the scale. Some behavior indicators suggested by the literature or the clinicians (e.g., hand perspiration) were not included in the final scale because pilot testing indicated that they were too difficult to observe accurately and reliably. Table 1 includes a description of each behavioral indicator of the final POSA, along with a reference to the article(s) supporting the use of each indicator in the scale.

Parent and teacher questionnaires: Independent anxiety measures. Several independent measures of anxiety were obtained for the purpose of validating the POSA. One was a questionnaire, the Parent Anxiety Rating Scale (Doris, McIntyre, Kelsey, & Lehman, 1971), completed by each child's parents and comprised of six questions about the child's separation anxiety (PARSEP) and 19 questions about the child's general anxiety (PARGEN). Two weeks before school began, this questionnaire was sent to the children's parents, who completed it by the beginning of the nursery school year. A second independent measure of anxiety was the Teachers' Separation Anxiety Scale (TSAS; Doris et al., 1971), comprised of 11 items about the child's reaction to separation from his/her mother or father when left at the nursery school at the beginning of the school day. One of two teachers (one for each of two nursery school sessions) rated the child on each of the first 10 consecutive days of the child's attendance at nursery school for the year. Ratings were made at the end of each school day and pertained to the period from the child's arrival at the nursery school with parent to the parent's departure. If the child was not delivered to the nursery school by his/her parent, the child was not rated that day, and his/her TSAS score was prorated.

Experimental manipulation of anxiety. In addition to obtaining questionnaire scores of the children's anxiety, we used a manipulation designed to create differing levels of anxiety in two experimental sessions, both involving cognitive tests. The first session was expected to be more anxiety arousing, since it involved an unfamiliar adult who individually tested each child in an unfamiliar setting during one of the first few days of the nursery school year. In contrast, the second session occurred approximately 2 weeks following the first session, after the children had had a chance to settle in to their new surroundings; furthermore, the child's mother was present at the second session, along with the same examiner who had tested the child at the first session. Finally, the children were given the same three tests in the second session that they had already taken in the first session.

Tests were given because it was felt that an evaluative atmosphere would make the experimental

situation more anxiety arousing, especially in the first session when the materials would be unfamiliar. Also, performance on the cognitive tests by the high-anxious children compared to the low-anxious children was of interest, since existing evidence is in conflict over this question. Some evidence suggests that high anxiety should interfere with test performance, especially when the tests are fairly difficult ones for the subjects (e.g., Feldhusen & Klausmeier, 1962; Tamaroff, 1976; Young & Brown, 1973). However, other evidence (e.g., Denny, 1966; Hodges & Durham, 1972; Katahn, 1966; Spielberger, 1966) suggests that anxiety may have facilitating effects on performance for children of higher socioeconomic and intellectual levels. (See other conflicting evidence in Fischer & Awrey, 1973; Mazzei & Goulet, 1969.) Since, according to the evidence, anxiety might lead to performance enhancement in some children and performance decrements in other children, no prediction was advanced regarding the relation between anxiety and task performance in the present study. However, the relation between task performance and anxiety scores was reported in an effort to shed light on the controversy just described.

Global anxiety self-ratings and ratings by the examiner. In the Introduction to this article, we emphasized the disadvantage of both self-reports by children and global (and thus subjective) ratings by adults. To determine whether our negative assessment was correct, we included two such measures in the present study, so that their effectiveness might be contrasted with that of the more specific and presumably more objective POSA. At the end of each of the two testing sessions, the examiner rated the child's general anxiety level during the session on a scale of 1 to 6. Also, a teacher asked each child to choose one of six pictures depicting progressively more fearful facial expressions to show how anxious the child felt during the testing session. The teacher rather than the examiner asked this question, since it was felt that the children would be more candid about their feelings with their teacher. The wording of the question was the following:

These pictures show a picture of a child who is more and more scared as you go from this top picture to the bottom picture here [point]. You see, up here the child is not scared at all [point]; here [second picture from top] the child is a little more scared but still pretty happy; here [third from top], he's getting more scared; and here [bottom], very, very scared. Which picture would show how scared you felt when you went downstairs to the testing room with that lady? One of these up here where the child isn't scared at all, or one of the bottom ones showing a child who is more and more scared?

The observation periods. At the beginning of the first session, the examiner, a 26-year-old experienced female teacher, approached each child individually in the nursery school and told the child

Table 1

Items of the Preschool Observation Scale of Anxiety and Interrater Agreement During the First 10 Minutes of the First Session

Item	Range of frequency	Interrater reliability*	Concordance 1	Concordance 2
1. Physical complaint: Child says he or she has a headache, stomachache, or has to go to the bathroom (B,C,D,E,G,K,Q)	0-0	nc ^b	nc	nc
2. Desire to leave: Child says he or she wants to leave the testing room or makes excuses about why he or she must leave; desire or "need" to leave must be explicit (D,E,P,Q).	0-5	.99*	.99	.46
3. Expression of fear or worry: Child complains about being afraid of or worried about something; must use the word "afraid," "scared," "worried," or a synonym (F).	0-0	nc	nc	nc
4. Cry: Tears should be visible (G,I,K,P,Q).	0-2	nc	nc	nc
5. Scream (P).	0-0	nc	nc	nc
6. Whine or whimper (G,P).	0-2	.69*	.97	.17
7. Trembling voice (F,G,I,M).	0-1	nc	nc	nc
8. Stutter (A,F,G,H,M,O).	0-0	nc	nc	nc
9. Whisper: Child speaks softly, without vocal cords; should not be a playful whisper (E,G).	0-11	.67*	.87	.39
10. Silence to one question in the interval (E).	0-3	.74*	.98	.50
11. Silence to more than one question in the interval (E).	0-5	.82*	.95	.22
12. Nail-biting: Child actually bites his or her nails in the testing room (F,G,I).	0-3	.83*	.98	.46
13. Lip-licking: Tongue should be visible (G).	0-13	.94*	.89	.63
14. Fingers touching mouth area: not counted if bites nails while touching mouth.	0-17	.96*	.91	.79
15. Sucking or chewing object: not fingernails (G,P).	0-1	.47*	.99	.33
16. Lip contortions.	0-13	.67*	.76	.44
17. Trembling lip (B).	0-0	nc	nc	nc
18. Gratuitous hand movement at ear area (G,I,J,N,P).	0-3	.56*	.96	.39
19. Gratuitous hand movement at top of head (G,I,J,N,P).	0-3	.62*	.95	.30
20. Gratuitous hand movement at an object separate from body or at a part of clothing separate from body (G,I,J,N,P).	0-18	.81*	.76	.61
21. Gratuitous hand movement at some part of body (not ear, hair, mouth, or genitals) (G,I,J,N,P).	0-12	.82*	.75	.56
22. Gratuitous hand movement (N).	0-9	.52*	.80	.35
23. Gratuitous leg movement (M, N).	0-20	.89*	.81	.77
24. Gratuitous foot movement: below ankles, distinguish from foot merely moving along with leg (M,N).	0-20	.49*	.66	.56
25. Trunk contortions (e.g., arching back) (N).	0-20	.89*	.81	.68

Table 1 (continued)

Item	Range of frequency	Interrater reliability ^a	Concordance 1	Concordance 2
26. Rigid posture: Part of body is held unusually stiff or motionless for the entire 30-sec interval (B,G,N).	0-11	.77*	.94	.16
27. Masturbation: touches genital area (K).	0-3	.81*	.99	.50
28. Fearful facial expression (E,I).	0-4	.92*	.98	.26
29. Distraction: Must be indicated by a verbal reminder by the examiner to the child to pay attention (B,I,L,P).	0-7	.86*	.94	.38
30. Avoidance of eye contact: Examiner should be having clear trouble making eye contact with child (G, M).	0-1	-.04 ^a	.98	.00
Total	26-107	.78*	.92	.58

Note. Letters in parentheses following each item description refer to studies suggesting that item as an anxiety indicator. The following code was used: A: Boland (1953); B: Buss, Wiener, Durkee, and Baer (1955); C: Cowen, Zax, Klein, Izzo, and Trost (1965); D: Endler, Hunt, and Rosenstein (1962); E: Fink (1956); F: Grossman (1968); G: Insel and Spencer (1972); H: Kasl and Mahl (1965); I: McReynolds (1965); J: Melamed and Siegel (1975); K: Miller, Barrett, Hampe, and Noble (1971); L: Nottelman (1975); M: Paul (1966); N: Raskin (1962); O: Santostefano (1960); P: Tamaroff (1976); Q: Wolff (1969). Items 14 and 16 were suggested by our clinician consultants. The range of frequency column gives the lowest score obtained by any subject and the highest score obtained by any subject. The interrater reliability column gives the correlation coefficients for the Pearson correlations between the scores of the two observers. The Concordance 2 column gives the interrater agreements on the occurrence of indicators in pairs of adjoining intervals (the number of agreements divided by the sum of agreements plus disagreements). The Concordance 1 column gives the interrater agreement on both the occurrence and nonoccurrence of indicators in pairs of adjoining intervals (again, the number of agreements by the sum of agreements plus disagreements).

^a n = 33 for interrater reliability and concordance calculations.

^b nc = not calculated due to infrequent occurrence (i.e., no children or only one child had nonzero scores on the indicator).

^c Based on only two children with nonzero scores.

* $p < .001$.

that she wanted him or her to go with her to another room "to do some tests."¹

The testing room was 5.7 m × 3.5 m, with one-way mirrors on top of its four walls and a microphone about 30 cm above the child's head for transmitting sounds from the testing room to the observation areas. A male and a female observer, both in their early 30s, sat behind the one-way mirrors at a distance of about 1.2 m from the child. The observers sat around a corner from each other, and were separated by about 1.5 m and a partition. Each observer spoke softly into a tape recorder each time he or she observed the child emit one of the 30 behavioral indicators. The two observers could not hear each other speaking. Also, the observers were blind to the independent anxiety ratings of the children whom they observed and to the fact that lower anxiety ratings were expected for the second session in comparison to the first session.

A standard time-sampling procedure involving 30-sec intervals was used. A Davis Scientific Instruments General Purpose Time Interval Generator (Model B501) emitted to each observer a beep and a red-light flash at the 30-sec intervals. Aside from

preventing observers from habituating to the behavioral indicators, this procedure was used to aid data analysis: Behaviors were given a score of 1 for each interval in which they occurred.

During the testing session, the examiner gave each child the three tests (Digits, Blocks, and Sentences) in standard form and in as neutral a manner as possible. The examiner was instructed to try to keep the children in the testing room for at least 10 minutes. The Digits test was taken from the Illinois Test of Psycholinguistic Abilities (Kirk, McCarthy, & Kirk, 1968) and involved repeating a series of orally presented digits from memory. Sentences were taken from Tamaroff's (1976) adaptation of the Sentences subtest of the Wechsler Preschool and Primary Scale of Intelligence (Wechsler, 1967). This test requires the child to repeat increasingly complex orally presented sentences from memory. Finally, Blocks was taken from Tamaroff's (1976)

¹ Two children refused to do any tests in the first session, and one refused to do any in the second session.

adaptation of the Block Construction Test of the Yale Scale of Child Development. It requires the child to copy visible block constructions with separate blocks, and it is timed.

In the second testing session, the same procedure was followed, except for two variations. First, the child was accompanied to the testing room by his or her mother as well as by the examiner, and the mother sat unobtrusively reading a magazine in the corner of the testing room during the session. Second, only the female observer was behind the one-way mirrors, since the second observer was used just during the first session to obtain interrater reliabilities. Of the 36 children in Session 1, 32 participated in Session 2. One child refused, and the parents of 3 children were unable to participate. Technical difficulties with recording equipment further reduced the sample (for whom complete data on both sessions were available) to 29.

Training for the observers. Both observers spent 2-3 hours memorizing the behavioral indicators and 6 hours in observation-training sessions. In training, the observers watched videotapes of three children being tested as other children would be tested in the actual experiment. Training also involved pilot observations of three children in the actual setting and conditions of the experiment. Finally, the observers spent approximately 2 hours discussing disagreements in their observations and ways to achieve better consensus. The detailed description of the items of the POSA given in Table 1 includes all of the details that the observers devised in order to maximize agreement.

Results

Scores for all subjects on the POSA were calculated by using the number of 30-sec intervals in which a given indicator occurred during the first 10 min (20 intervals) of the sessions. The examiner had been instructed to try to keep each subject in the testing room and working for a full 10 min; thus, a 10-min interval was chosen as the target time, in part, because of concern that the experimenter's behavior might have changed significantly after the 10-min period had expired.²

Interrater Reliabilities

To assess interrater reliability, Pearson correlations between the ratings of the two observers were calculated using the Session 1 scores for each indicator separately and for the sum of all 30 indicators (see Table 1 third column from the right). The most im-

portant correlation coefficient, that for the 30 indicators together, was .78 ($p < .001$). The intraclass correlation was .77 ($p < .001$).

Even though the preceding analysis answered the central question about reliability of overall POSA and individual indicator scores, we also sought to learn the level of specificity at which observer agreement took place. Toward this end we used a demanding procedure designed to gauge the degree of concordance between observers *within* observation intervals. The procedure involved grouping every 2 adjoining intervals (resulting in 19 interval groups) and counting how often the observers agreed or disagreed as to whether an indicator occurred within each interval group. Adjoining rather than single intervals were used for this agreement measure, since the observers sometimes reported the same behavior at slightly different times so that the interval cutoff occurred between their reports. Using this method of measurement, a quotient of concordance was calculated by dividing the number of agreements between the observers by the number of their "agreements" plus "disagreements" for each indicator separately and for the total of all 30 indicators. Table 1 (last two columns) shows the results of these calculations (a) when agreements between observers that a given indicator did *not* occur were included in the "agreement" score and (b) when such negative agreements were excluded from the agreement score. Considering the rigorous nature of the procedure (particularly b above), the percentages shown in Table 1 represent rather substantial agreement between observers at the level of brief observation intervals.

² The alternate scoring method of dividing the frequency of intervals in which each indicator occurred during the total session by the number of intervals in the total session would have posed an additional problem, in that fatigue and familiarity could have affected the scores of children having longer sessions. Unfortunately, using the 10-minute cutoff, all subjects were not engaged in the same activities during the target time, but this seemed to be a less significant consideration than those noted above.

Table 2
Correlations Among Anxiety Measures and Test Scores

	2	3	4	5	6	7	8	Sentences
1. POSA	.37*	.30*	.47**	-.12	-.02	-.04	.38*	-.16
2. PARSEP		.41**	.36*	.01	.06	.05	.28	.24
3. PARGEN			.22	-.18	.07	.03	.05	.05
4. TSAS				-.05	.19	-.07	.07	-.18
5. Self-rating					.14	-.45**	-.35*	-.24
6. Examiner rating						-.24*	-.10	-.13
7. Blocks							.42**	.62***
8. Digits								.72***

Note. POSA = Preschool Observation Scale of Anxiety; PARSEP = questions about the child's separation anxiety; PARGEN = questions about the child's general anxiety; TSAS = Teachers' Separation Anxiety Scale.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Correlations With Independent Measures of Anxiety

As an initial step in assessing the validity of the POSA, the correlations of the POSA with the PARSEP, PARGEN, and TSAS were calculated using data from the first session (since correlations with data of the second session would have been less meaningful due to the mothers' presence). As Table 2 indicates, all three correlations were significant. Other correlations noted in the table are also of interest. Note that the self-rating and examiner ratings were not correlated with one another or with POSA, PARSEP, PARGEN, or TSAS scores. Thus, even though the POSA did meet the first set of validity criteria that we established (i.e., significant correlations with the three inventory measures), the self-ratings and examiner ratings did not. The significant negative correlations between self-ratings and two of the cognitive tests suggests that self-ratings may have been influenced by the children's awareness of the quality of their test performance.

To assess the contributions of individual POSA indicators to the correlations between the scale and the three inventory measures, the correlations of each of the 30 items with the three inventory scores were calculated. Most of these individual correlations were nonsignificant, suggesting that the predictive power of the indicators lies mainly in their combination with one another. Those corre-

lations that did attain statistical significance were lip contortions with PARSEP ($.39, p < .01$) and TSAS ($.43, p < .01$), gratuitous hand movement in the ear area with PARGEN ($.32, p < .05$), gratuitous hand movement at the top of the head with PARSEP ($-.38, p < .01$), gratuitous hand movement toward object with TSAS ($.53, p < .001$), gratuitous hand movement at other body part ($-.31, p < .05$), gratuitous leg movement with PARSEP ($.34, p < .05$) and TSAS ($.35, p < .05$), gratuitous foot movement with PARSEP ($.35, p < .05$) and TSAS ($.46, p < .01$), trunk contortions with PARSEP ($.29, p < .05$) and TSAS ($.42, p < .01$), and masturbation with PARGEN ($.39, p < .01$) and TSAS ($.58, p < .001$).

Anxiety Scores in the First and Second Sessions

As a second way of assessing the validity of the POSA, the mean POSA score for all children in the first session was compared with the mean score for all children in the second session (designed to be less anxiety producing than the first). As predicted, the children obtained significantly higher POSA scores in the first than in the second session, $t(28) = 2.53, p < .01$ (one-tailed). Of the 29 children, 22 had higher scores in the first than in the second session, $\chi^2(1) = 6.76, p < .01$. Further comparisons between the first and second sessions were made for children

Table 3
Comparisons of POSA Scores Between Sessions

Group	Session 1		Session 2		<i>t</i>	<i>df</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
All children	61.4	21.3	51.3	14.0	2.53*	28
High PARSEP scorers	70.0	19.6	49.7	15.5	4.35**	14
High PARGEN scorers	64.3	19.2	49.0	14.3	3.17*	15
High TSAS scorers	61.4	22.4	47.6	14.0	2.66*	17

Note. POSA = Preschool Observation Scale of Anxiety; PARSEP = questions about the child's separation anxiety; PARGEN = questions about the child's general anxiety; TSAS = Teachers' Separation Anxiety Scale.

* $p < .01$, one-tailed.

** $p < .001$, one-tailed.

who scored above the median on the PARSEP, PARGEN, and TSAS, since it was thought that these children might be especially sensitive to the situational manipulations of stressors. Again, the high scorers on the three questionnaire scales showed significantly more behavioral indicators of anxiety in the first than in the second session. The results for these analyses are shown in Table 3. Note that in each of the three groups, the magnitude of the Session 1–Session 2 difference is greater than for the entire sample. In fact, of the four groups, children rated by their parents as high in separation anxiety showed the highest mean POSA scores during Session 1 with mothers absent and the largest Session 1–Session 2 difference—more than doubling the difference shown by the full sample.

Session 1–Session 2 differences were also calculated for the children's self-ratings and the examiner ratings. The change in the children's self-ratings from Session 1 to Session 2 indicates a nonsignificant increase in anxiety, whereas the change in the examiners' ratings indicates a highly significant decrease. Anxiety ratings on the children's 6-point picture scale averaged 2.22 in Session 1 and 2.59 in Session 2 ($p = .28$). Ratings on the examiner's 6-point scale averaged 3.07 in Session 1 and 2.00 in Session 2, $t(29) = 3.71$, $p < .001$. Thus, both the POSA and examiner ratings of anxiety met the second validity criterion, that is, significantly higher scores in Session 1 than in Session 2.

Finally, Session 1–Session 2 differences for each of the 30 POSA indicators were

calculated in an effort to gauge the contribution of the individual indicators to the overall sessions difference in total POSA scores. As was true in the correlational analysis reported earlier, most individual item effects were nonsignificant, suggesting that the discriminative power of the indicators lies principally in their combination with one another. However, there were five indicators that had significantly different frequencies in Sessions 1 and 2 (all in the predicted direction): silence to one question ($p < .05$), touching mouth area ($p < .01$), gratuitous arm movement ($p < .001$), trunk contortions ($p < .001$), and rigid posture ($p < .05$).

Discussion

The results of the present study support the use of the POSA as a measure of anxiety in young children. Independent judges achieved strong agreement on both total POSA scores and total scores for most of the 30 individual indicators. The large number of indicators used seemed to interfere with interobserver concordance at the micro level of 1-min observation blocks (at least by the most rigorous method of analysis); and this suggests that the two observers may have differed frequently in the particular behavioral incidents that they observed. Yet, this is a relatively trivial limitation considering the high interobserver correlations obtained for total POSA and individual indicator scores. In sum, the difficulties introduced by requiring observers to watch for

30 indicators appear to be outweighed by the value of including a variety of potential anxiety manifestations in order to capture indications of anxiety in individuals with differing expressive styles.

Two types of evidence support the view that the POSA yields a valid index of anxiety. As predicted, the scale was significantly correlated with all three inventory measures of anxiety. Also as predicted, the POSA yielded significantly higher scores during a presumably high-anxiety test session than during a session designed to provoke less anxiety. The data generally supported our original belief that broad-based measures such as the POSA would outperform simpler approaches such as self-ratings by young children and global judgments by examiners. Children's self-ratings and examiner ratings did not correlate significantly with any of the three inventory measures, with POSA scores, or with each other. Self-ratings showed very slight differences (and in the wrong direction) between Sessions 1 and 2; however, Session 2 ratings by the examiner were significantly lower than Session 1 ratings. It is uncertain whether this latter finding derived from an expectation by the examiner that children's anxiety would be lower in the second session with mothers present. However, whatever the basis for the finding, it constitutes the only bit of evidence supporting the use of either self- or examiner ratings. This pattern of findings seems to indicate the superiority of structured observations of carefully delineated behaviors over global, unstructured, and thus subjective ratings, though this conclusion must be qualified by the fact that the structured observations were made by trained observers, whereas the global ratings were not.

As indicated in the Introduction, the relation between anxiety and problem-solving performance appears to be quite complex. In the present study none of the inventory measures was significantly related to test performance, and the POSA was significantly related only to Digits performance ($r = .38$). So, anxiety, as measured by the POSA and inventory scores, did not appear to have a debilitating effect on test performance for

children in the present study, and may have had an enhancing effect on one test, perhaps due to factors related to the high socioeconomic status of the present subjects.

Further research with the POSA should include investigations of the scale's capacity to reflect the effects of stressors other than those devised in the present study. Children varying more widely in age and other demographic characteristics than those of the present sample should be included. And, more importantly, there is a need to assess the scale's usefulness with clinical populations of children who (unlike those of the present sample) suffer from pronounced behavior problems. In addition, research using longer observation periods than the 10-minute-plus sessions of the present study could be useful; some behaviors that did not occur in the relatively brief sessions of the present investigation might prove to be useful indices of anxiety if children were given a lengthier opportunity to display them. Even the behaviors that occur infrequently might well prove to be potent anxiety indicators when they do appear. Finally, the degree of anxiety is apt to be reflected not only by the *frequency* of anxiety behaviors but also by their *intensity*. Although intensity may be difficult to quantify, its potential for increasing the precision of anxiety measurement would seem to justify efforts in this direction.

For the present, however, the POSA represents a potentially useful approach to the assessment of anxiety in children. Given the sensitivity of the POSA to situational variations in stressors, the scale itself should be particularly useful to researchers concerned with the interplay between specific situational conditions and affective states (e.g., Spielberger et al., 1972). In addition, the POSA could contribute usefully to our understanding of sex differences in anxiety manifestations at various ages (see Maccoby & Jacklin, 1974, pp. 182-190), and to our capacity to evaluate therapeutic techniques for children (see Achenbach, 1974, pp. 606-650). Research on these topics, focusing on anxiety states in children, clearly must circumvent the problems of projective techniques, self-reports, global and subjective

observer reports, and physiological measures, all outlined in the Introduction. The evidence presented in the present study suggests that instruments such as the POSA may provide logical, valid, and reliable alternatives to these more traditional approaches.

Reference Note

1. Hollingshead, A. B. *Two-factor index of social position*. Unpublished manuscript, Yale University, 1957.

References

- Achenbach, T. M. *Developmental psychopathology*. New York: Ronald Press, 1974.
- Boland, J. L. A comparison of stutterers and non-stutterers on several measures of anxiety. *Speech Monographs*, 1953, 20, 144.
- Buss, A., Wiener, M., Durkee, A., & Baer, M. The measurement of anxiety in clinical situations. *Journal of Consulting Psychology*, 1955, 19, 125-129.
- Cowen, E. L., Zax, M., Klein, R., Izzo, L. D., & Trost, M. A. The relation of anxiety in school children to school record, achievement, and behavioral measures. *Child Development*, 1965, 36, 685-695.
- Denny, J. P. Effects of anxiety and intelligence on concept formation. *Journal of Experimental Psychology*, 1966, 72, 596-602.
- Doris, J., McIntyre, A., Kelsey, C., & Lehman, E. Separation anxiety in nursery school children. *Proceedings of the 79th Annual Convention of the American Psychological Association*, 1971, 79, 145-146. (Summary)
- Endler, N. S., Hunt, J. M., & Rosenstein, A. J. An S-R Inventory of Anxiousness. *Psychological Monographs*, 1962, 76(17, Whole No. 536).
- Feldhusen, J. F., & Klausmeyer, H. J. Anxiety, intelligence, and achievement in children of low, average, and high intelligence. *Child Development*, 1962, 33, 403-409.
- Fink, R. M. How children tell us they are afraid. *Childhood Education*, 1956, 32, 264-266.
- Fischer, D. G., & Awrey, A. Manifest anxiety, test anxiety, and intelligence in concept formation. *Journal of Social Psychology*, 1973, 89, 153-154.
- Grossman, B. D. Anxiety as a factor in the child's responsiveness to social reinforcement. *Journal of Genetic Psychology*, 1968, 112, 165-174.
- Hodges, W. F., & Durham, R. L. Anxiety, ability, and digit span performance. *Journal of Personality and Social Psychology*, 1972, 24, 401-406.
- Insel, S. A., & Spencer, T. D. *Emotional concomitants of reading readiness thinking in kindergarten*. Washington, D. C.: National Center for Educational Research and Development, December 1972. (ERIC Document Reproduction Service No. ED 072 871)
- Kasl, S. V., & Mahl, G. F. The relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology*, 1965, 1, 425-433.
- Katahn, M. Interaction of anxiety and ability in complex learning situations. *Journal of Personality and Social Psychology*, 1966, 3, 475-479.
- Kirk, S. A., McCarthy, J. J., & Kirk, W. D. *Illinois Test of Psycholinguistic Abilities*. Urbana: University of Illinois Press, 1968.
- Lazarus, R. S. *Psychological stress and the coping process*. New York: McGraw-Hill, 1966.
- Maccoby, E. E., & Jacklin, C. N. *The psychology of sex differences*. Stanford, Calif.: Stanford University Press, 1974.
- Mazzei, J., & Goulet, L. R. Test anxiety and intelligence in concept formation. *Psychological Reports*, 1969, 24, 842.
- McReynolds, P. On the assessment of anxiety: I. by a behavior checklist. *Psychological Reports*, 1965, 16, 805-808.
- McReynolds, P. The assessment of anxiety: A survey of available techniques. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 1). Palo Alto, Calif.: Science and Behavior Books, 1968.
- Melamed, B. G., & Siegel, L. J. Reduction of anxiety in children facing hospitalization and surgery by use of filmed modeling. *Journal of Consulting and Clinical Psychology*, 1975, 43, 511-521.
- Miller, L. C., Barrett, C. L., Hampe, E., & Noble, H. Revised anxiety scales for the Louisville Behavior Check List. *Psychological Reports*, 1971, 29, 503-511.
- Nottelman, E. *Test anxiety and off-task behavior in evaluative situations*. Washington, D.C.: National Institute of Education, April 1975. (ERIC Document Reproduction Service No. ED 113 022)
- Paul, G. L. *Insight vs. desensitization in psychotherapy*. Stanford, Calif.: Stanford University Press, 1966.
- Phillips, B. N., Martin, R., & Meyers, J. Interventions in relation to anxiety in school. In C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (Vol. 2). New York: Academic Press, 1972.
- Raskin, A. Observable signs of anxiety or distress during psychotherapy. *Journal of Consulting Psychology*, 1962, 26, 389.
- Santostefano, S. Anxiety and hostility in stuttering. *Journal of Speech and Hearing Research*, 1960, 3, 337-347.
- Sarason, S. B. The Measurement of anxiety in children: Some questions and problems. In C. D. Spielberger (Ed.), *Anxiety and behavior*. New York: Academic Press, 1966.
- Sarason, S. B., Davidson, K. S., Lighthall, F. F., & Waite, R. R. Rorschach behavior and the performance of high and low anxious children. *Child Development*, 1958, 29, 277-285.
- Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., & Ruebush, B. K. *Anxiety in elementary school children*. New York: Wiley, 1960.

- Spielberger, C. D. The effects of anxiety on complex learning and academic achievement. In C. D. Spielberger (Ed.), *Anxiety and behavior*. New York: Academic Press, 1966.
- Spielberger, C. D. Anxiety as an emotional state. In C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (Vol. 1). New York: Academic Press, 1972.
- Spielberger, C. D., et al. *Needed research on stress and anxiety. A special report of the OSOE-sponsored grant study: Critical appraisal of research in the personality-emotions-motivation domain* (IBR Report No. 72-10). Washington, D.C.: Office of Education, July 1972. (ERIC Document Reproduction Service No. ED 113 649)
- Tamaroff, M. H. Separation anxiety and cognitive functioning in nursery school children. (Doctoral dissertation, Cornell University, 1975) *Dissertation Abstracts International*, 1976, 36, 5820-B. (University Microfilms No. 76-9569)
- Wechsler, D. *Wechsler Preschool and Primary Scale of Intelligence*. New York: Psychological Corporation, 1967.
- Weisz, J. R. Transcontextual validity in developmental research. *Child Development*, 1978, 49, 1-12.
- Wolff, S. *Children under stress*. London: Penguin Press, 1969.
- Young, F. A., & Brown, M. Effects of test anxiety and testing conditions on intelligence test scores of elementary school boys and girls. *Psychological Reports*, 1973, 32, 643-649.

Received August 17, 1977 ■

Diagnostic Validity of a Standardized Neuropsychological Battery Derived from Luria's Neuropsychological Tests

Charles J. Golden

University of Nebraska Medical Center

Thomas A. Hammeke and Arnold D. Purisch

University of South Dakota

The work of A. R. Luria has been recognized as a major contribution to neuropsychology. Among his accomplishments, Luria has devised an extensive set of procedures used for neuropsychological evaluation. Luria's tests permit the full identification of the specific deficits underlying a disorder and can be completed in about 2 hours. The most significant flaw in the battery is a lack of standard administration and scoring that has precluded an assessment of its validity. The present study was an attempt to overcome these deficiencies by developing an objective form, combining Luria's procedures with the advantages of a standard test battery. The resultant test was evaluated using 50 medical control and 50 neurological patients. Of the 285 measures in the battery, 253 significantly discriminated at the .05 level, and only 16 failed to discriminate at the .2 level. A discriminant analysis, using the 30 most effective items, yielded a hit rate of 100%. The battery's potential and the future research necessary are discussed.

The work of A. R. Luria, the Russian neuropsychologist, has been internationally recognized as a major contribution to experimental and clinical neuropsychology. Luria has made extensive theoretical contributions (e.g., Luria, 1963, 1966, 1970, 1973) and has devised numerous clinical diagnostic and rehabilitation procedures (e.g., Luria, 1963, 1966). Many of Luria's diagnostic procedures have recently been published by Christensen (1975a, 1975b) along with the equipment necessary to perform these tests (Christensen, 1975c).

Luria's tests are based on his theoretical contributions to neuropsychology. Luria envisions the brain divided into three principal units responsible for arousal, sensory input and integration, and behavioral planning and execution. In turn, the areas within each of

these units have specific functions. For example, one area within the second functional unit is responsible for auditory input, and another area is responsible for the integration of visual and auditory stimuli.

Within Luria's system, overt behavior is the result of the cooperation among different areas of the brain. The pattern of interacting areas responsible for a given behavior is called a functional system. Each area of the brain participates in numerous functional systems.

The effect of any brain injury is to interrupt the execution of any functional system that includes the injured area(s). Thus, any brain injury will effect the performance of numerous behaviors. The generality of the behavioral loss depends on the importance of the functional systems interrupted, as well as the availability of alternate functional systems to replace the injured system.

The role of the neurodiagnostician is to isolate the specific point at which a client's functional systems have been interrupted. This requires a detailed evaluation of a sub-

Requests for reprints or for additional information on the battery should be sent to Charles J. Golden, Nebraska Psychiatric Institute, University of Nebraska Medical Center, Omaha, Nebraska 68105.

ject's intact and interrupted functional systems. From the pattern of performance deficit discovered, the specific locus of the brain injury can be determined, as well as such factors as type and severity of the disorder.

Thus, Luria's diagnostic tests consist of numerous specific procedures. These are designed to isolate dysfunction, compared to the more global assessments characteristic of many neuropsychological tests. In addition, Luria's procedures tend to be qualitative in nature, rather than quantitative.

Luria's procedures have several practical advantages. First, they provide a more extensive breakdown of behavior than can be determined from more global tests. Thus, the clinician can derive a more specific analysis of the deficits present in a patient. This information can be used for both neurodiagnosis and for specific rehabilitation planning based on an individual's pattern of test results (Golden, 1976, 1978; Luria, 1963).

A second advantage of these test procedures is the ability to do a full examination in less than 2.5 hours (Luria, 1966). In contrast, a test battery like the Halstead-Reitan, which can reveal similar deficit patterns, may take from 6 to 8 hours. A third advantage of the Luria tests is their portability. Luria's procedures were specifically set up to be done at bedside with a minimum of inexpensive equipment. Again, this is in contrast to many current test procedures that require a laboratory setting or great amounts of equipment, sometimes costing several hundred dollars.

Finally, the Luria battery includes an evaluation of all of the areas necessary for a complete neuropsychological exam (Benton, 1975). This includes evaluation of motor functions, sensory skills (auditory, tactile, and visual), verbal skills (expressive speech, receptive speech, reading, and writing), spatial skills, mathematical abilities, memory, and intellectual skills.

Despite these advantages, the Luria tests have been widely criticized. One significant problem has been the lack of any direct evaluation of the tests beyond Luria's own observations and conclusions. In this regard,

Reitan (1976) has written that Luria's "reports essentially represent evaluation of 'critical cases' based upon his own observations, conclusions, and statements of significance" (p. 199). This lack of systematic validation has greatly limited the generalizability of Luria's results.

A major problem in a validation study is the lack of any *standardized* method of giving or scoring Luria's test procedures. In many cases, the description of administrative procedures is vague. Moreover, Luria acknowledges changing procedures for individual patients. Even when administrative procedures are clear, scoring procedures are not. Scoring is determined by the personal assessment of the clinician based on experience and knowledge rather than on any normative data.

Despite these major problems, the Luria battery clearly possesses a high degree of face validity as well as a strong theoretical foundation. With proper standardization, scoring, and validation, the Luria tests could become a major tool of both clinical and experimental neuropsychology.

Development and Rationale

The intention of the present study was to use the material presented by Luria (1966, 1973) and Christensen (1975a, 1975b, 1975c) to form a standardized, objectively scored version of Luria's neuropsychological procedures. The new battery was designed to retain, as closely as possible, the qualitative nature of Luria's tests and the sampling of all the major areas of neuropsychological performance.

At the same time, the standardization of the items and the objective scoring would allow for careful validation and replication of the test results, as well as for the systematic collection of data on a wide variety of neurological disorders. This would allow the development of scales representing specific loci of injury or underlying causes in order to establish a firmer interpretation of the battery.

Consequently, the battery, as envisioned, would have the advantages of a detailed,

qualitative analysis of a client's behavior, as well as the advantages of a standardized quantitative battery and the systematic diagnostic research that can be completed with such a battery. In addition, the battery would have the advantage of a relatively short testing time.

The first step in designing the battery was to develop standardized items, keeping as close as possible to the original procedures while creating items that could be objectively scored. After the development of the initial items, they were administered to a clinical population to evaluate their practicality.

As the project progressed, numerous items were rewritten or deleted as they were found to be ineffective or duplicative. Others were eventually discarded when a reliable scoring system could not be developed for them. Some items were eventually found to be too difficult to administer to an impaired population. After 6 months of effort, a final version of the test to be used in this study was developed.

Present Study

The purpose of the present study was an initial validation of the standardized battery. Fifty normal hospitalized subjects were compared with 50 brain-injured subjects. *t* tests were calculated on each score within the battery. In addition, a discriminant analysis was used to measure the overall effectiveness of the battery.

Method

Subjects

The subjects were 100 hospitalized patients tested in hospitals in Sioux City, Iowa; Yankton, South Dakota; and Sioux Falls, South Dakota. Fifty subjects had confirmed neurological diagnosis made on the basis of medical exams alone by a qualified physician, usually a neurologist or neurosurgeon. The control subjects had a variety of medical problems including back injuries, infectious diseases, and chronic cases of pain.

The average age of the control subjects was 42.0 years ($SD = 14.8$ years), and the average age of the experimental subjects was 44.3 years ($SD = 18.8$ years). The difference in age was not significant,

$t(98) = .7$, $p > .40$. Overall, there were 49 females and 51 males. No significant differences in sex distribution were present in the two groups.

The control group had 12.21 years ($SD = 2.86$ years) of education, and the neurological group had 10.30 years ($SD = 2.84$ years). The difference between the groups was significant, $t(98) = 3.51$, $p < .01$.

Test Battery

Overall, the Luria-South Dakota Neuropsychological Test Battery consists of 285 measures and can be administered in less than 2½ hours to a significantly impaired individual. The battery requires several pieces of inexpensive equipment. First, a series of cards with pictures and word items published by Christensen (1975c) are needed. Several additional pictures (available from the first author) are necessary to replace some of the Christensen items that were found to be ineffective with an American population.

In addition, the battery requires the following objects: (a) a 13-cm black comb; (b) a 2½ × ½ inch (5.5 × .6 cm) rubberband; (c) a paper clip, jumbo size; (d) a Bow Compass 5178 (available from the Empire Pencil Company, Shelby, Tennessee 37160); (e) a Pedigree Quality Eraser 2910 (also available from the Empire Pencil Company); (f) a key (WR2 Curtis 177); (g) a straight pin; (h) a quarter; (i) a metric ruler; and (j) an audiotape for some of the rhythm and verbal items.

The items in the battery are generally adapted from Christensen (1975a, 1975b, 1975c)² with slight modifications as necessary in order to establish a standard administrative or scoring procedure. The items fall roughly into 10 categories:

Motor functions. This section includes a series of tasks requiring the reproduction of simple motor movements with the hands, mouth, and tongue, both when a model is provided and under verbal instructions alone. The section also evaluates simple coordination, optical-spatial organization, complex sequencing of behavior, and the ability to draw. Sample items on this scale include:

1. Using your right hand, touch your fingers in turn with your thumb as quickly as you can

¹ Information on obtaining the materials used in the battery, copies of the battery, and the current test manual can be obtained from the first author.

² The items in the battery have been modified from Luria's Neuropsychological Battery with permission. (Copyright 1975 by Anne-Lise Christensen and Munksgaard, Copenhagen, Denmark, and Spectrum Publications, Inc., 1975-20 Wexford Terrace, Jamaica, New York 11432.)

while you count them. [with palm facing up, demonstrate and then have the subject practice before timing. Allow 10 sec. Score the number of complete times the subject does the sequence accurately.]

5. Close your eyes and place your right hand in the same position as I place it first. [Press thumb against fifth finger for 2 sec., then return to normal position. Score correct or incorrect.]

11. Do as I do. [Place right hand under chin with fingers bent.]

22. With your hands in front of you, tap your right hand twice and your left hand once, changing smoothly from one hand to the other like this. [Demonstrate and allow the subject to practice.] Do this as quickly as you can until I tell you to stop. [Allow 10 sec. The score is the number of fully correct sequences.]

25. Show me how to work with scissors.

36. Without lifting your pencil from the paper, I want you to draw the best square you can. [Score for time and quality. All quality items have objective grading requirements in the test manual.]

44. [Show picture of a square]. Draw this. [Score as above].

50. If I knock once, raise your right hand, if I knock twice, raise your left hand. [Give four trials, alternating one and two knocks. Score number of errors.]

Rhythm (acoustico-motor functions). This section includes items requiring the individual to differentiate between sounds with different pitch and rhythmic relationships. The subject must indicate whether sounds are the same or different as well as whether they reproduce rhythmic and pitch patterns. Sample items included:

52. Now you are going to hear two tones on a tape. Tell me whether the tones are the same or different. [Play six pairs of tones. The score is the number of errors.]

53. Again, you will hear two tones. Tell me which is higher, the first or second tone. [Play tape with five pairs. Score errors.]

58. Tell me how many beeps you hear. [Play four groups of beeps from tape. Score number of errors.]

62. You will now hear a rhythm on the tape. When I tell you that the rhythm is over, I want you to tap with your hand the rhythm you heard on the tape. [Play three rhythms. Score as right or wrong.]

Cutaneous and kinesthetic functions (tactile). This section evaluates complex cutaneous function, muscle and joint sensations, and stereognosis. Kinesthetic assessment requires a blindfolded subject to identify the direction of limb movements and reproduce limb positions. Cutaneous assessment includes evaluation of threshold, localization, stimulus identification, and two-point finger discrimination. The assessment for stereognosis requires a blindfolded subject to identify common objects placed in the palm of the hand under both active and passive palpating conditions. Sample items include:

64. Tell me where I am touching you. [Have blindfolded subject in sitting position with hands in front and palms facing up. Touch the subject with the eraser end of pencil, alternating among right and left fingers (numbered 1-5), palms (P), forearms (F), and shoulders (S). If uncertain of where the subjects report touch from verbal report, have the subjects indicate the place touched with their opposite hand. Touches should be in the following order:

Right hand: 1 F 3 5 P 2 S 4
Left hand: P 2 3 S 5 4 F 1

Score for number of errors on each hand.

80. Now I will put your left arm in a certain position; try to put the other arm in the same position. [Extend left arm of blindfolded subject in front at 90°.]

82. Feel this object and tell me exactly what it is. [Instruct the subject to hold their right palm up and place objects on the fingers. Alternate between hands in this manner. Allow 10 sec per item. Score for correctness of answer, and time each response. Objects include quarter, key, eraser, and jumbo paper clip.]

Visual functions. This section includes a series of tasks assessing the integrity of visual-spatial perception, including the identification of objects and pictures, identifying the missing elements in complex geometric configurations (similar to the tasks in Raven's Progressive Matrices) and constructing geometric patterns from blocks. Subjects must identify time on clocks with no numbers and show spatial and directional orientation. Finally, the ability of a subject to perform spatial rotations and transformations is assessed.

Sample items include:

86. What do you call this object? [Examiner presents the subject with objects one at a time. Allow 10 sec per item. Items are pencil, eraser, rubber band, and quarter.]

87. What is this picture supposed to be? [Present pictures one at a time and allow 10 sec for each.

Pictures include a purse, a nut cracker, a glass vial, a camera, and an egg carton. Score the number wrong.]

88. [Show clock faces.] Tell me what time these clock faces show.

97. This drawing shows a stack of blocks in three dimensions. Tell me how many blocks are in the stack. [Be sure to include those you see as well as those you don't see.]

Impressive speech. This section assesses a subject's ability to discriminate basic English phonemes and to reproduce the discriminations orally or by writing; to name familiar and unfamiliar objects among a series of pictures; and to respond to statements and questions that require the understanding of genitive, prepositional, comparative, and complex grammatical constructions.

Sample items included:

100. You will hear some sounds on the tape. What I want you to do is first repeat exactly the sound you hear and then write down the letter of the alphabet that goes along with the sound. For example, if you hear the sound "ta," you would say "ta" and then write down the letter "t". [Score oral and writing errors separately.]

110. I will place some pictures before you. I want you to point at the shoe, the candle, the stove.

114. Put your hand on your head.

125. Which boy is shorter if John is taller than Peter?

130. Is the following sentence said by a disciplined or an undisciplined person? "I am unaccustomed to disobeying rules."

Expressive speech. This section includes tasks requiring the articulation of simple speech sounds, familiar and unfamiliar words of varying lengths, and phrases or sentences of varied length and complexity. The test also requires the subject to name and classify objects and to produce narrative descriptions.

Sample items include:

133. Repeat after me: (a) a (as in late); (b) i (as in light); (c) m (as in milk) (d) b (as in baby); and (e) sh (as in shine).

137. Repeat after me: (a) hairbrush, screwdriver, laborious.

138. Repeat after me: house-ball-chair. [Say all three as one item].

157. What objects do these pictures represent? [Show five pictures.]

163. Say the days of the week backwards starting with Sunday.

164. Tell me what's happening in this picture. [Present picture. Score time to respond and number of words in first five seconds of response.]

Reading and writing. This section requires the subject to break words into their component sounds or letters; to integrate sounds or letters into words; to copy letters and words; to write words of varying complexity from dictation; and to read sounds, words, phrases and paragraphs.

Sample items:

178. Copy these in your own handwriting. [Present card.]

183. Write these words [dictate]: wren, knife.

188. What sound is made by the letters (a) g-r-o and (b) p-l-y.

192. Read these sounds [Present card with the sounds po, cor, cra, spro, and prot.]

197. Read these sentences. [Present cards.]

Arithmetical skills. This section requires a subject to identify Arabic and Roman numerals, to identify the significance of digit placement, to compare numbers of varying size, and to do simple arithmetic operations (multiplication, addition, subtraction) and simple algebraic manipulations. The ability to form arithmetic series is also evaluated. Items are presented both orally and visually.

Sample items:

203. Write these numbers: 71, 17, 69, 96.

206. Read these numbers [on card]: 7-9-3, 3-5-7.

210. Tell me which is larger: 17 or 68? 23 or 56? 189 or 201?

216. Add these numbers in your head: 5, 9, 7.

219. What is the missing number [Present card with the following:]

$$12 - \underline{\quad} = 8.$$

$$12 + \underline{\quad} = 19.$$

Mnemonic processes. This section involves a series of tasks assessing an individual's retention and retrieval skills for visual, acoustic, and kinesthetic inputs. Subjects must work with both verbal and nonverbal material. The effects of retroactive and proactive interference are also examined.

Sample items:

227. I am going to show you a card, and I want you to look at it carefully. When I remove the card, I want you to draw as much from it as you can remember. [Present card for 7 sec.]

231. I want you to remember some words that I am going to say: house-tree-cat. Now, look at this picture; what do you see? [Allow the subject to describe the picture for 15 sec.] What were the words?

Intellectual processes. This section requires the subject to interpret the themes of pictures, to demonstrate vocabulary skills, to form concepts, to classify objects, to understand analogies, to complete complex arithmetical problems, and to show logical reasoning skills.

Sample items:

245. What is meant by the expression "iron hand"? "green thumb"? [Score for quality.]

249. In what way are table and sofa alike? In what way are an ax and a saw alike?

256. What has the same relation to good as high has to low?

258. Peter had two apples and John had six apples. How many did they have together?

Other items in the test, in addition to the examples shown, test the same processes, varying the difficulty, modality involved, instructions, or hand that is used.³

Procedure

In the case of each potential subject, written permission was obtained to review medical records in order to assess whether the subject met the necessary requirements of a confirmed diagnosis for the study. Individuals whose diagnosis was questionable as to the presence or absence of brain injury were not included in the study, unless a definitive diagnosis was later established.

Once a patient had assented to participate, administration was arranged at a time that did not interfere with the hospital schedule. The majority of test administrations occurred at bedside, although free rooms for testing were occasionally available. Because of the relative short time needed to administer the battery, the testing was done in a single session. In some cases, however, the session was divided by visitors or hospital tests.

Scoring

Items were scored by a variety of methods according to the nature of the item and the quali-

tative dimension or dimensions that were to be assessed by the item. As can be seen in the sample items already presented, many items could be scored as right or wrong, or the overall number of errors could be counted on an objective basis. Time to complete an item or latency of response was also frequently used. Other scoring criteria included frequency of response in a specified time, trials to correct performance, and number of items completed. In some cases, quality of response was judged. The test manual gives extensive instructions on the evaluation of such responses.

For all items, the raw scores were recorded and then converted into a 0, 1, 2 system. A score of 0 was intended to represent the performance characteristic of a normal individual. A score of 1 represented performance intermediate to that of normals and brain-damaged patients. A score of 2 represented performance characteristic of brain damage. The scoring for each item was established by finding cutoff points that maximized the discriminative effectiveness of the item in a group of 75 subjects collected early in the development of the battery (37 normals, 38 brain damaged). This was done by maximizing the chi-square that compared the number of subjects in each category who were diagnosed correctly or incorrectly.

Alternate scoring systems were also tried that contained wider ranges, including a 0, 1, 2, 3 system and a 0, 1, 2, 3, 4 system. However, these failed to add any additional discriminative validity to the tests. Hence, it appears that the simpler system should be used. For researchers wishing to compare scores on tests more finely, the raw score data allow this to be accomplished.

To evaluate the reliability of the scoring system, the test was administered by 1 examiner in the presence of a second examiner. Each examiner scored the test independently. This procedure was repeated for five patients. In each repetition, a different pair of examiners was used; thus the procedure involved 10 examiners. Overall, there were 1,425 pairs of scores available. The rate of agreement was over 95%.

Results

t tests were run between the control and neurological group on all 285 scores generated by the test. Of the 285 comparisons made, 253 were significant at the .05 level ($df = 98$). In all cases, the neurological group performed less effectively (a higher score) than the control group. Of the 32 items not significant at the .05 level, 16 exceeded the .20 level of probability. On these

³ Copies of the text, a test manual describing scoring criteria, and procedures for each item are available from the first author.

32 items, the neurological group performed poorly on 30 items and showed identical performance on two items.

Because of the educational differences between the groups, one-way analyses of variance were calculated using education as a covariate. In no case did any significant item become nonsignificant because of the control of the educational difference, although education contributed a small amount of variance to each item. In no case did a nonsignificant item become significant because of the use of the covariate.

To further evaluate the potential effectiveness of the battery, a discriminant analysis was run using all 285 measures. It was found that a weighted, linear combination of 30 variables was sufficient to separate the groups with 100% accuracy.

Discussion

The results of the current study clearly supports the validity of the standardized battery. Nearly 90% of the items suggested by Luria, and modified so as to be objective, significantly discriminated between the brain-injured and normal subjects. In addition to the excellent results with the individual items, the discriminative analysis achieved a 100% hit rate using only 30 of the 285 scores. This result is comparable to the results achieved by a neuropsychological test battery of any kind. It is recognized that the hit rate might not be as high in a cross-validation. Despite this, the results clearly illustrate the strong potential of the assessment approach reflected in the battery.

An evaluation of those items that failed to discriminate between groups revealed no consistent pattern. Some of the items were too easy and were missed by no one, such as "Show me your teeth." Others were clearly quite difficult: "Which is correct, 'The earth is illuminated by the sun' or 'The sun illuminates the earth'?" (Correct answer: Both are correct). In other cases, no fully comprehensive scoring method could be devised. In this category, one deceptively simple item "Show me how to frown" resisted repeated attempts to devise an effective scoring procedure.

Despite the impressive results of the current study, there is a need for more extensive research before the battery can be used in clinical situations on a regular basis. One important research goal will be the establishment of a basic summary scoring system so that performance in general areas (spatial skills, speech, cognition, among others) can be easily determined and compared.

A second research area is the demonstration of the effectiveness of the battery in a population including psychiatric patients, a setting in which extensive neuropsychological evaluation takes place. It is necessary to examine the effectiveness of the battery with psychiatric patients who have a chronic history and a long duration of hospitalization, subjects for whom neuropsychological tests are often highly ineffective (Golden, 1977, 1978; Lezak, 1976).

A third research area is assessing the ability of the battery to localize injuries and aid in the identification of underlying neurological processes. Luria's theoretical system suggests that the battery should be highly effective in this regard, perhaps more effective than any other comparable neuropsychological battery.

In this regard, it may also be possible to develop scales designed to specifically measure such factors as laterality, localization, or process. This scale development would be analogous to the process used with the Minnesota Multiphasic Personality Inventory and other similar tests.

The final area of necessary research is to examine the effects of variables such as age, education, medication, intelligence, chronicity, and severity of a disorder. As the effects of these variables are determined, appropriate clinical procedures can be devised to correct for these factors or to include them in a clinical analysis.

In addition to its usefulness in neurodiagnosis, the battery is a potential useful instrument in rehabilitation planning. As was seen in the Method section, the battery evaluates a wide range of abilities, allowing the clinician to identify the specific areas in which an individual has problems as well as the underlying deficits in an individual's functional systems caused by a brain injury.

There is also the ability to evaluate those abilities that remain intact. This information can be used to plan rehabilitation tasks designed to retrain a lost ability or to use intact abilities to reformulate a functional system (Golden, 1976, 1978; Luria, 1963). This potential of the battery requires careful and systematic evaluation by rehabilitation psychology.

At present, we are carrying out further studies aimed at meeting some of the important research requirements outlined above. The present study clearly indicates that the standardized Luria battery has tremendous potential and may be the extensive, highly effective, economical, and standardized battery that neuropsychology will need as the field grows into more diverse settings. Future research should determine the extent to which the battery fulfills this potential.

References

- Benton, A. L. Psychological tests for brain damage. In H. Freedman, H. Kaplan, & B. Sadock (Eds.), *Comprehensive textbook of psychiatry* (Vol. 2). Baltimore: Williams & Wilkins, 1975.
- Christensen, A. L. *Luria's neuropsychological investigation*. New York: Spectrum, 1975. (a)
- Christensen, A. L. *Luria's neuropsychological investigation: Manual*. New York: Spectrum, 1975. (b)
- Christensen, A. L. *Luria's neuropsychological investigation: Text cards*. New York: Spectrum, 1975. (c)
- Golden, C. J. The role of the psychologist in the training of the neurologically impaired. *Professional Psychology*, 1976, 7, 579-584.
- Golden, C. J. Validity of the Halstead-Reitan Neuropsychological Battery in a mixed psychiatric and brain-injured population. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1043-1051.
- Golden, C. J. *Diagnosis and rehabilitation in clinical neuropsychology*. Springfield, Ill.: Charles C Thomas, 1978.
- Lezak, M. *Neuropsychological assessment*. New York: Oxford University Press, 1976.
- Luria, A. R. *Restoration of function after brain damage*. New York: Pergamon Press, 1963.
- Luria, A. R. *Higher cortical functions in man*. New York: Basic Books, 1966.
- Luria, A. R. *Traumatic aphasia*. Paris: Mouton, 1970.
- Luria, A. R. *The working brain*. New York: Basic Books, 1973.
- Reitan, R. M. Neurological and physiological basis of psychopathology. *Annual Review of Psychology*, 1976, 27, 189-216.

Received August 18, 1977 ■

Discrimination of Schizophrenic and Brain-Injured Patients by a Standardized Version of Luria's Neuropsychological Tests

Arnold D. Purisch
University of South Dakota

Charles J. Golden
University of Nebraska Medical Center

Thomas A. Hammeke
University of South Dakota

This study examines the ability of a standardized battery of tests suggested by the extensive work of A. R. Luria to discriminate between brain-injured and schizophrenic patients. An earlier study reported 93% effectiveness for the standardized battery in discriminating brain-injured patients and normal controls. In the present study, the battery was administered to 100 schizophrenic and brain-injured patients. Chronicity was 121 months ($SD = 125$) for the schizophrenic group and 56 months ($SD = 116$) for the neurological patients. Of the 282 items in the battery, schizophrenics showed significantly better performance on 72 items ($p < .05$). A discriminant analysis using 60 items demonstrated 100% diagnostic accuracy. Schizophrenics performed better on 10 of 14 summary measures ($p < .01$). A discriminant analysis using the 14 summary measures achieved 88% diagnostic accuracy. The accuracy shown by the battery is as high as the results obtained using other tests in a comparable chronic schizophrenic sample. The usefulness of the battery in schizophrenic and brain-injured patients is discussed.

The differential diagnosis between psychological problems due to schizophrenia and those due to brain injury is a significant issue in the field of clinical neuropsychology. Several diagnostic instruments have been developed for this purpose, ranging from short, individual test procedures such as the Bender-Gestalt to long test batteries such as the Halstead-Reitan.

Despite many attempts to find the ideal test or tests, no instrument yet developed consistently separates schizophrenic from brain-injured groups. Although many studies have reported statistically significant group differences (Yates, 1954, 1966), these results do not reflect the real clinical utility of a test. A test can discriminate between groups on the basis of mean differences in performance, but there may still be too much over-

lap between members within each group to allow accurate individual diagnosis. Therefore, the major measure of clinical effectiveness is the hit rate—the percentage of patients accurately diagnosed by a test (Yates, 1954).

Unfortunately, many studies do not report hit rates (Spren & Benton, 1965). Studies that have reported hit rates when evaluating the diagnostic effectiveness of an assessment technique in discriminating between schizophrenic and brain-injured patients have indicated a wide scope of results ranging from chance levels of 50% to over 90% accuracy. Generally, no test or test battery has shown consistently good results in studies using psychiatric patients (see Golden, 1978).

Three primary factors have been identified by various reviewers to explain the inconsistent results in the literature: differences in chronicity of schizophrenia, the comprehensiveness of assessment procedures, and the methods of data analysis (Davison, 1974). In general, chronicity is positively related to

Requests for reprints and copies of the test and test manual should be sent to Charles J. Golden, Nebraska Psychiatric Institute, University of Nebraska Medical Center, Omaha, Nebraska 68105.

decrements in adaptive functioning as measured by a wide variety of psychometric tests. Long-term chronic schizophrenics are likely to reveal deficits similar to those seen in neurological patients. The majority of studies reporting nonsignificant results or low hit rates have used long-term chronic schizophrenics. For example, Watson and his collaborators (Watson, 1968, 1972; Watson, Thomas, Andersen, & Felling, 1968; Watson, Thomas, Felling, & Andersen, 1968; Watson, Thomas, Felling, & Andersen, 1969; Watson & Ueker, 1966) tested schizophrenics who had an average length of hospitalization of over 10 years. Those studies reporting positive results have used schizophrenics with generally short hospitalizations of under 1 year and less severe disorders (e.g., Golden, 1977).

A second factor is the ability of the tests to assess all neuropsychological skills. Organicity cannot be considered as a unitary entity resulting in identical disturbances across patients (Boll, 1974). In a mixed brain-injured group, the differences in the nature and loci of lesions for each individual result in unique patterns of disturbed functioning. Procedures not providing comprehensive assessment may fail to reveal lesions that manifest disturbances in abilities other than those included in the assessment. Accordingly, the Halstead-Reitan, presently the most comprehensive standardized test battery, has been able to demonstrate the highest hit rates with neurologically impaired populations (Golden, 1977).

Another problem exists with some of the commonly used tests of brain damage. These tests are comprised of items that require the concerted working of several neuropsychological abilities (Lezak, 1976). These types of tasks are often so complex that even non-organic subjects, particularly schizophrenics, may perform poorly. As a result, many patients without neurological problems may be falsely diagnosed as organically impaired. For example, poor performance on the Bender-Gestalt may not be a positive indicator of brain damage; it may simply reveal psychological dysfunction without regard to etiology (Golden, 1977).

The method of analyzing test results is a

third factor to consider in evaluating the efficacy of various diagnostic instruments. Most studies of tests that do not demonstrate significant group differences have based their comparisons on global measures representing overall test performance. Performance on individual test items is obscured in these global measures, and much potentially useful diagnostic information is lost. As such, global quantitative comparisons may not be a fair method of assessing a test's true diagnostic effectiveness (Lezak, 1976). For example, Goldstein and Neuringer (1966) argued that a qualitative analysis of the Trail Making Test, part of the Halstead-Reitan, increased hit rates to nearly 80%, whereas global scores were only able to achieve separation at chance levels. Hewson (1949a, 1949b) demonstrated up to 90% diagnostic accuracy using analyses of Wechsler Adult Intelligence Scale (WAIS) subtest relationships, whereas studies comparing group mean performances did not show consistent significant differences (Matarazzo, 1972).

The statistical methods that are used may also affect the results of a study. Many studies have used univariate tests to detect group differences for each measure under consideration. This practice may result in lower hit rates than when maximum information is obtained by combining tests. A study by Golden (1978) demonstrated more reliable discrimination between schizophrenics and brain-injured subjects when the results of all measures in a test battery were statistically combined by a discriminant analysis than when each measure was considered separately by *t* tests.

A. R. Luria, the Russian neuropsychologist, has developed a set of procedures intended to assess systematically each area of neuropsychological functioning. Based on Luria's extensive theoretical contributions to neuropsychology (e.g., Luria, 1966, 1973), the items yield a qualitative assessment of the major neuropsychological skills.

Luria's procedures have several advantages. First, they allow for a comprehensive assessment of neuropsychological abilities. Second, complex behaviors are systematically analyzed by items assessing specific neuro-

psychological skills. This feature provides information about the qualitative nature and degree of a disturbance that can be useful for diagnostic and rehabilitative purposes (Luria, 1963). Finally, a relatively short administration time makes it more likely that a patient's attention and interest can be sustained (Golden, 1978; Smith, 1975). These characteristics suggest that it should be a highly useful diagnostic instrument, even in chronic schizophrenic populations.

Despite these advantages, the tests have been criticized. Stressing the importance of flexibility when testing neurological patients, Luria does not provide a standard manner in which to administer his tests. Instead, he recommends modifying procedures as necessary to account for patient needs and specific referral questions. In addition, Luria does not provide for rating responses. Judgment and clinical intuition are the methods of test interpretation rather than comparisons to normative standards. Reitan (1976) has criticized the lack of standardization and the lack of scoring criteria, contending that Luria's opinion is the only measure of validity that has been used with the battery.

As a result of these criticisms, we (Golden, Hammeke, & Purisch, 1978) developed an alternate version of Luria's battery. We attempted to eliminate the weaknesses dis-

cussed above while maintaining the positive features. Standard administration procedures were defined, and a scoring system was developed. The scoring included separate scores for the important dimensions of each response. Another consideration was to keep administration time as short as possible without sacrificing comprehensiveness.

Hammeke, Golden, and Purisch (in press) has shown the standardized battery to be highly effective in discriminating brain-injured and hospitalized control patients. We found that the individual items in the test discriminated experimental and control patients 100% of the time. We also found that a combination of 14 summary scores could achieve a 93% hit rate. These rates are as good as or better than the results seen in any other single test or test battery (e.g., see Golden, 1977; Lezak, 1976).

Since the discrimination of brain-damaged and schizophrenic patients appears to be the most difficult task for a neuropsychological test battery or procedure, the present study is an attempt to evaluate the effectiveness of the standardized version of Luria's tests when comparing these two populations. If the test is able to make this discrimination, then the clinical value of the Luria in diagnostic work would be strongly increased.

Method

Subjects

The subjects were 100 hospitalized patients. All were approached at the recommendation of their physician, and all volunteered to participate in the study. Fifty subjects had confirmed neurological diagnoses made on the basis of medical examination by a qualified physician, usually a neurologist or neurosurgeon. The 50 other subjects were patients diagnosed by psychiatrists as having a schizophrenic disorder. No schizophrenics were included for whom there was a possibility of organicity as indicated by a history of seizures, alcoholism, or head trauma. Table 1 presents the diagnostic subtypes of both the schizophrenic and brain-injured subjects.

Overall, there were 56 male and 44 female subjects. No significant difference in the sexual distribution of the two groups was present. Other background data were obtained from the subjects' hospital files after securing their signed informed consent. Table 2 presents the means and standard deviations of each group for age, education, length of hospitalization, duration of illness (chronicity), age of onset of illness, and number of previous hospitalizations.

Table 1
Frequency and Percentage of Diagnostic Subtypes for Brain-Injured and Schizophrenic Subjects

Diagnostic subtype	f	%
Neurological diagnosis		
Cerebral trauma	10	20
Neoplasms	6	12
Infectious diseases	3	6
Cerebral vascular disorder	14	28
Degenerative diseases	6	12
Epilepsy	4	8
Metabolic and toxic disorders	3	6
Congenital disorders	4	8
Schizophrenic diagnosis		
Catatonic	2	4
Hebephrenic	3	6
Paranoid	19	38
Simple	1	2
Undifferentiated	20	40
Schizoaffective	5	10

Table 2
Means, Standard Deviations, and *t* Tests for Demographic Indices

Index	Brain injured ^a		Schizophrenic ^a		<i>t</i> ^b
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Age (years)	44.36	18.83	41.32	14.52	.90
Education (years)	10.30	2.84	11.38	2.59	1.97
Length of hospitalization (days)	19.62	43.38	410.42	1,569.32	1.76
Chronicity (months)	56.24	116.40	121.02	125.03	2.67**
Age of onset (years)	39.42	22.34	32.08	12.26	2.04*
Previous hospitalizations	1.50	2.05	4.29	3.48	4.79***

^a*n* = 50.

^b*df* = 98.

**p* < .05.

***p* < .01.

****p* < .0001.

This table indicates that schizophrenics had significantly higher means for chronicity and number of previous hospitalizations. Schizophrenics also had a significantly earlier mean age of onset compared to the brain-damaged subjects. Length of present hospitalization was considerably longer for the schizophrenics, but this difference was not significant. No significant group differences were found for age or years of education.

Test Battery

All of the items in the test battery were adapted from Christensen (1975a, 1975b, 1975c) with modification in order to establish standard administrative scoring procedures. The battery consists of 282 measures and can be administered in less than 2½ hours. The items fell into 10 categories.¹

Motor functions. This series of tasks requires the reproduction of simple motor movements with the hands, mouth, and tongue both when a model is provided and from verbal instructions alone. The section also evaluates simple coordinated abilities, motor-spatial organization, complex sequencing of behavior, and the ability to draw.

Acoustico-motor (rhythm) functions. These items require the individual to differentiate between sounds with different pitch and different rhythmic relationships. The subject must indicate whether sounds are the same or different and reproduce given pitch or rhythmic relationships played from a tape.

Cutaneous and kinesthetic (tactile) functions. This section evaluates complex cutaneous functions, muscle and joint sensations, and stereognosis. Kinesthetic assessment requires a blindfolded subject to identify the direction of limb movements and reproduce limb position. Cutaneous assessment includes evaluation of threshold localization, stimulus identification, and two-point finger and palm discrimination. The assessment for stereognosis requires a blindfolded subject to identify common objects placed in the palm of the hand.

Visual functions. These tasks assess the integrity of visual-spatial perception, including the identification of objects and pictures, identifying the missing elements in complex geometric configurations (similar to the tasks in Raven's Progressive Matrices), transposing pictures of blocks with no numbers, and showing spatial and directional orientation. Finally, the ability of a subject to perform spatial rotations and transformations is assessed.

Expressive speech. This section includes tasks requiring the articulation of simple speech sounds, familiar and unfamiliar words of varying lengths, and phrases or sentences of varied length and complexity. The test also requires the subject to name and classify objects and to produce narrative descriptions.

Impressive speech. The tasks require the articulation of simple speech sounds, familiar and unfamiliar words of varying lengths, and phrases or sentences of varied length and complexity. The subject must name and classify objects and produce narrative descriptions.

Reading and writing. The subject must break words into their component sounds or letters; integrate sounds or letters into words; write words of varying complexity from dictation; and read letters, words, phrases, and paragraphs.

Arithmetic skills. The subject is required to identify Arabic and Roman numerals, to identify the significance of digit placement, to compare numbers of varying size, and to perform simple arithmetic operations (multiplication, addition, subtraction) and simple algebraic manipulations. The ability to form arithmetic series is also evaluated. Items are presented both orally and visually.

¹ Due to time limitations and the desire to allow each article to be independently read, the present article reproduces some material from the previous article (Golden, Hammeke, & Purisch, 1978) on the standardized Luria test battery.

Mnemonic processes. The tasks assess an individual's retention and retrieval skills for visual, acoustic, and kinesthetic inputs. Items involve both verbal and nonverbal material. The effects of retroactive and proactive interference are also examined.

Intellectual processes. The final section requires the subject to interpret the themes of pictures, to demonstrate vocabulary skills, to form concepts, to classify objects, to understand analogies, to understand complex arithmetical problems, and to show logical reasoning skills.

The battery requires several pieces of inexpensive equipment. First, a series of cards with pictures and word items published by Christensen (1975c) are needed. Several additional pictures are necessary to replace some of the Christensen items that were found to be confusing for an American population. In addition, the battery requires (a) a 13-cm black comb; (b) a $2\frac{1}{4} \times \frac{1}{4}$ inch ($5.5 \times .6$ cm) rubberband; (c) a paper clip, jumbo size; (d) a Box Compass 5178 (available from Empire Pencil Company, Shelby, Tennessee 37160); (e) a Pedigree Quality Eraser 2910 (also available from the Empire Pencil Company); (f) a key (WR2 Curtis 177); (g) a straight pin; (h) a quarter; (i) a metric ruler; (j) an audiotape for some of the rhythm and verbal items; and (k) a blindfold.

Procedure

In the case of each potential subject, written permission was obtained to review medical records in order to assess whether the subject met the requirements of a confirmed diagnosis for the study. An individual whose diagnosis was questionable as to the presence or absence of brain injury was not included unless a definitive diagnosis was later established.

Once a patient had consented to participate, administration was arranged at a time that did not interfere with the hospital schedule. The majority of the test administrations occurred at bedside, although free rooms for testing were occasionally available. Because of the relatively short time needed to administer the battery, most of the testing was done in a single session. In some cases, however, the session was interrupted by visitors or medical tests.

Scoring. Items were scored by a variety of methods according to the nature of the item and the qualitative dimension that was being assessed. These scoring methods included accuracy of response, frequency of response, adequacy of response, number of errors, time for performance, trials to correct performance, and number of items completed. For many items, more than one scoring method was used to measure different performance dimensions.

All raw scores were recoded into a 3-point (0, 1, 2) scaled score. A scaled score of 0 was intended to represent the performance characteristic of a normal individual. A scaled score of 1 represented

an intermediate level of performance seen in both brain-injured and normal individuals. A scaled score of 2 represented the performance characteristic of brain-injured subjects. The scoring for each item was established by examining the performance of a mixed diagnostic group of 75 subjects collected in the first stages of this project.

Summary scoring indices. The scores of the items in each section were summed to yield a summary index for that function. Ten summary indices were created in this manner.

Four additional scoring indices were also created. The first measure summed the 34 most effective indicators of brain damage. These were items that when scored as 2, were nearly always indicative of brain damage. These items were selected on the basis of the performance of a mixed neurological, schizophrenic, and normal population collected during the development of the battery (Golden et al., 1978). This index was labeled the *pathognomic* index.

The second index was the sum of all items that required right-hand motor or tactile/kinesthetic function. The third index was similarly devised for all left-hand items. These indices were labeled *left hemisphere* and *right hemisphere*, respectively.

The final index represented overall performance. The scores on each of the 13 previous summary indices were converted into *z* scores using the means and standard deviations of data previously collected on a hospitalized normal control group. These 13 *z* scores were then summed into a final index, labeled a *total* score.

Results

Two-tailed *t* tests were computed for the scaled scores on all 282 measures between the schizophrenic and brain-injured groups. Of the 282 comparisons, the schizophrenics performed significantly better on 72 items at the .05 level ($df = 98$). The brain-injured group demonstrated better performance on 2 items at the .05 level of significance ($df = 98$). Using a stepwise discriminant analysis of the individual items, it was found that a hit rate of 100% could be achieved with 40 items.

The means and standard deviations were calculated for each of the 14 summary indices. Differences between the two groups were then determined by the two-tailed *t* tests, as reported in Table 3. The schizophrenics performed significantly better than the brain-injured subjects at the .01 level on 9 of the indices and at the .001 level on 1 other index. Four indices failed to significantly discriminate between the groups.

Table 3
Means, Standard Deviations, and *t* Tests for Summary Indices

Index	Brain injured ^a		Schizophrenic ^a		<i>t</i> ^b
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Motor	44.36	19.70	34.20	17.53	2.73*
Rhythm	12.21	5.59	12.26	5.70	-.05
Tactile	19.32	12.00	13.62	8.58	2.73*
Visual	16.14	5.53	12.82	5.32	3.06*
Impressive speech	23.67	12.82	19.48	10.49	1.79
Expressive speech	33.34	16.59	24.86	14.34	2.73*
Reading and writing	24.67	11.91	17.51	10.51	3.19*
Arithmetic	16.15	12.43	10.07	9.21	2.78*
Memory	16.25	6.05	15.84	6.61	.33
Intellectual	38.32	15.53	33.24	13.80	1.73
Pathognomic	34.46	12.10	22.20	9.46	5.65**
Left hemisphere	18.82	10.91	13.24	8.37	2.87*
Right hemisphere	18.00	11.12	12.44	8.35	2.83*
Total	27.06	21.04	15.51	17.95	2.95*

^a *n* = 50.

^b *df* = 98.

* *p* < .01.

** *p* < .001.

Cross-tabulations were calculated to determine the optimal cutoff point for each of the 14 summary indices. The percentage of correct diagnoses for each index was determined using the cutoff point. A discriminate analy-

sis was calculated to determine the overall effectiveness of the summary indices. As can be seen in Table 4, even though none of the individual indices were able to discriminate at a hit rate of better than 74%, the combi-

Table 4
Percentage of Correct Classifications Yielded by Cutoff Value for the Summary Indices

Index	Cutoff ^a	% correctly classified		
		Brian injured ^b	Schizo-phrenic ^b	Total ^a
Motor	40	54	66	60
Rhythm	8	78	34	56
Tactile	14	66	66	66
Visual	14	68	66	67
Impressive speech	22	52	70	61
Expressive speech	26	70	68	69
Reading and writing	20	58	74	66
Arithmetic	9	64	64	64
Memory	16	62	54	58
Intellectual	35	58	64	61
Pathognomic	26	70	78	74
Left hemisphere	13	68	62	65
Right hemisphere	7	38	92	65
Total	13	74	56	65
Overall	—	84	92	88

^a Cutoff point was chosen to maximize percentage of total classifications. Subjects with a score less than or equal to the cutoff were classified as schizophrenic.

^b *n* = 50.

^c *n* = 100.

nation of indices was able to discriminate at an overall hit rate of 88%

Discussion

The results of this study demonstrate a high effectiveness for the standardized battery in discriminating between schizophrenia and brain injury. No other test or set of tests reported in the literature have achieved a comparable 100% hit rate. The 88% hit rate using the 14 summary indices is also higher than the results reported in any comparable studies. Most of the positive results reported for other tests were obtained using schizophrenic populations of slight to moderate chronicity compared to the long-term chronic schizophrenics used in this study.

Summary indices were devised to make the test easier to interpret with individual patients. Of the 14 indices, only the rhythm, impressive speech, memory, and intelligence summary measures did not significantly discriminate between the groups. The schizophrenic group performed in the brain-damaged range on these four scales, compared to a previously tested normal group (Golden et al., 1978). The acoustico-motor items required greater sustained concentration and attention than other items, which could have resulted in difficulties for schizophrenics, who found it hard to maintain this focus. The memory index included many interference tasks that may have triggered irrelevant interfering associations. The intellectual index, assessing higher abstract thinking and verbal-reasoning abilities, and the impressive speech index, which is disproportionately weighted with items presenting complex verbal relationships, were relatively difficult for the schizophrenics compared to the more basic indices.

On the other hand, schizophrenics demonstrated superior performance on the motor, tactile, visual, left-hemisphere, and right-hemisphere indices. Similarly, most of the items from other indices on which schizophrenics performed better than the brain-injured patients did not require complex symbolic manipulations, sustained attention, or higher abstract ability. These results are con-

sistent with the theoretical position that neurological damage results in impaired functioning for both simple and complex tasks, whereas cognitive impairment associated with schizophrenia results in greater disruption on tasks requiring complex verbal abilities than on tasks requiring little verbal mediation.

The hit rates obtained for each of the 14 indices demonstrate a 10%-20% decrease in diagnostic accuracy compared to the results obtained using hospitalized normal controls. It is clear that none of these indices used alone would be diagnostically accurate with long-term chronic schizophrenics. Nevertheless, these results are equal to or better than those found for other tests in similar populations. In addition, the 88% hit rate obtained is comparable to the 93% with the hospitalized normal controls (Hammeke et al., in press). This result underscores the necessity of considering all indices in determining differential diagnoses between long-term chronic schizophrenic and brain-injured patients.

The ability to comprehensively assess neuropsychological functioning in relatively little time is one of the major advantages of the battery. Yet, the development of effective summary measures from the rich item pool may further shorten administration time when there are specific diagnostic questions. For example, an index comprised of items chosen for their sensitivity to the general effects of brain damage could be developed as a simple screening device when there is a general question of presence or absence of brain damage. Indices used to answer other questions can be developed in a similar manner.

In addition to its clinical usefulness, the battery also shows good research potential. Unlike most other tests, the scoring of many items reflects a single qualitative dimension of a response, allowing component neuropsychological abilities to be more easily isolated and analyzed. Empirical relationships can be easily established between specific neuropsychological deficits associated with specific neurological disorders, with schizophrenia, and with such variables as medication and prognosis.

Overall, the results demonstrated that the standardized battery is able to make the difficult discrimination between long-term chronic schizophrenics and brain-injured patients. The battery's main advantages are short administration time, comprehensiveness, and a systematic breakdown of complex neuropsychological abilities. These features make the battery particularly well suited both for use as a research instrument in clarifying many of the presently unanswered questions about the underlying processes of schizophrenia and as a clinical tool for the neurodiagnostician working with schizophrenic and brain-injured patients.

References

- Boll, T. J. Psychological differentiation of patients with schizophrenia versus lateralized cardiovascular, neoplastic, or traumatic brain damage. *Journal of Abnormal Psychology*, 1974, 83, 456-458.
- Christensen, A. L. *Luria's neuropsychological investigation*. New York: Spectrum, 1975. (a)
- Christensen, A. L. *Luria's neuropsychological investigation: Manual*. New York: Spectrum, 1975. (b)
- Christensen, A. L. *Luria's neuropsychological investigation: Text cards*. New York: Spectrum, 1975. (c)
- Davison, L. A. Current status of clinical neuropsychology. In R. Reitan & L. Davison (Eds.), *Clinical neuropsychology: Current status and applications*. Washington, D.C.: V. H. Winston, 1974.
- Golden, C. J. Validity of the Halstead-Reitan Neuropsychological Battery in a mixed psychiatric and brain-injured population. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1043-1051.
- Golden, C. J. *Diagnosis and rehabilitation in clinical neuropsychology*. Springfield, Ill.: Charles C Thomas, 1978.
- Golden, C. J., Hammeke, T. A., & Purisch, A. D. Diagnostic validity of a standard neuropsychological battery derived from Luria's neuropsychological tests. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1258-1265.
- Goldstein, G., & Neuringer, C. Schizophrenic and organic signs on the Trail Making Test. *Perceptual and Motor Skills*, 1966, 22, 347.
- Hammeke, T. A., Golden, C. J., & Purisch, A. D. Validity of Luria's neuropsychological investigation in the diagnosis of cerebral dysfunction. *International Journal of Neuroscience*, in press.
- Hewson, L. R. The Wechsler-Bellevue Scale and the Substitution Test as aids in neuropsychiatric diagnosis: Part I: *Journal of Nervous and Mental Disease*, 1949, 109, 158-183. (a)
- Hewson, L. R. The Wechsler-Bellevue Scale and the Substitution Test as aids in neuropsychiatric diagnosis: Part II. *Journal of Nervous and Mental Disease*, 1949, 109, 246-260. (b)
- Lezak, M. D. *Neuropsychological assessment*. New York: Oxford University Press, 1976.
- Luria, A. R. *Restoration of function after brain injury*. New York: Macmillan, 1963.
- Luria, A. R. *Higher cortical functions in man*. New York: Basic Books, 1966.
- Luria, A. R. *The working brain: An introduction to neuropsychology*. New York: Basic Books, 1973.
- Matarazzo, J. D. *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). Baltimore: Williams & Wilkins, 1972.
- Reitan, R. M. Neurological and physiological basis of psychopathology. *Annual Review of Psychology*, 1976, 27, 189-216.
- Smith, A. Neuropsychological testing in neurological disorders. *Advances in Neurology*, 1975, 7, 49-110.
- Spren, O., & Benton, A. L. Comparative studies of some psychological tests for cerebral damage. *Journal of Nervous and Mental Disease*, 1965, 140, 323-333.
- Watson, C. G. The separation of NP organics from schizophrenics on three visual motor screening tests. *Journal of Clinical Psychology*, 1968, 24, 412-414.
- Watson, C. G. Cross-validation of a sign developed to separate brain damaged from schizophrenic patients. *Journal of Clinical Psychology*, 1972, 28, 66-67.
- Watson, C. G., Thomas, R. W., Andersen, D., & Felling, J. Differentiation of organics from schizophrenics at two chronicity levels by a use of the Reitan-Halstead Organic Test Battery. *Journal of Consulting and Clinical Psychology*, 1968, 32, 679-684.
- Watson, C. G., Thomas, R. W., Felling, J., & Andersen, D. Differentiation of organics from schizophrenics with Reitan's sensory-perceptual disturbances tests. *Perceptual and Motor Skills*, 1968, 26, 1191-1198.
- Watson, C. J., Thomas, R. W., Felling, J., & Andersen, D. Differentiation of organics from schizophrenics with the Trail Making, Dynamometer, Critical Flicker Fusion, and Light-Intensity Matching tests. *Journal of Clinical Psychology*, 1969, 25, 130-133.
- Watson, C. J., & Ueker, A. E. An attempted cross-validation of the Minnesota Percepto-Diagnostic test. *Journal of Consulting Psychology*, 1966, 30, 461.
- Yates, A. J. The validity of some psychological tests of brain damage. *Psychological Bulletin*, 1954, 51, 359-379.
- Yates, A. J. Psychological deficit. *Annual Review of Psychology*, 1966, 17, 111-114.

Self-Administered Relaxation Training and Money Deposits in the Treatment of Recurrent Anxiety

Clifford E. Lewis, Anthony Biglan, and Elizabeth Steinbock
University of Oregon

This study evaluated two self-administered relaxation manuals and a money deposit in the treatment of recurrent, nonphobic anxiety in a college population. Subjects were randomly assigned to a self-monitoring-only control group or one of four active treatment conditions. Subjects in active conditions received a progressive relaxation manual or a manual that called for the client to devise his or her own relaxation method and were assigned to deposit or nondeposit conditions. Improvement did not differ for the two relaxation procedures, but relaxation training groups improved significantly more than self-monitoring-only subjects on both self-report questionnaires and self-monitored measures of anxiety. The money deposit did not produce greater amounts of relaxation practice or adherence to the program, although subjects in the money deposit condition did report being more relaxed in practice sessions and improved more on two pre-post measures of anxiety. Subjects' locus of control scores were significantly related to a number of practice, adherence, and outcome variables, but subjects' ratings of the likelihood that they would practice and benefit from the program proved to be as good predictors. The study suggests the value of self-monitoring and relaxation practice as treatment for recurrent, nonphobic anxiety.

Although considerable progress has been made in our ability to treat phobias (cf. Bandura, 1969; Kazdin & Wilcoxon, 1976; Paul, 1969), we have not yet identified an effective strategy for treating nonphobic, anxious clients. Despite the large number of phobic treatment studies, there have been surprisingly few systematic investigations of how we might effectively assist clients who

experience recurrent anxiety for which no stimuli can be identified or for which stimuli are too numerous to allow the use of phobic treatment procedures. In this study, we evaluated two forms of self-administered relaxation training as treatments for recurring, nonphobic anxiety in a college population and examined whether a money deposit increases the number of practice sessions that clients complete.

This study was supported by Grant MH27559-01 from the National Institute of Mental Health. It is based on a doctoral dissertation conducted by the first author, was supervised by Gerald Kranzler, and was submitted to the Department of Counseling of the University of Oregon.

Clifford Lewis is now at the Department of Rehabilitation, State of California.

Elizabeth Steinbock is now at Developmental Evaluation Clinic, Duke University Medical Center, Durham, North Carolina.

We appreciate the assistance of Richard Haller and Jennifer Scheetz.

Requests for reprints should be sent to Anthony Biglan, Department of Psychology, University of Oregon, Eugene, Oregon 97403.

Perhaps the most promising approach to nonphobic anxiety involves the use of some form of relaxation training. There is evidence from case studies that relaxation can benefit chronically anxious persons (Dawley, 1975; Raskin, Johnson, & Rondestvedt, 1973). Sherman's work (Sherman, 1975; Sherman & Plummer, 1973) suggests that persons can be taught to continue to use relaxation in stressful situations. Both Zeisset (1968) and Goldfried and Trier (1974) have presented evidence that relaxing as an active "coping skill" is beneficial at least for situationally specific anxieties. Thus, a program of relaxa-

tion training with instructions to relax in problematic situations could be an effective strategy for treating persons who suffer from recurrent, nonphobic anxiety.

It may not be necessary, however, to use the progressive or deep muscle relaxation procedure in such a program. It is possible that any procedure that helps clients interrupt chains of anxious responses will benefit them. Numerous responses other than deep muscle relaxation have been shown to reduce anxiety, including oriental defense exercises (Gershman & Stedman, 1971), muscle tension (Wolpin & Raines, 1966), and meditative procedures (Benson, Beary, & Carol, 1974). Kass, Rogers, and Feldman (1973) reported success in six of seven desensitization cases in which clients used anxiety-inhibiting responses that were already in their repertoires. This approach could be better than progressive relaxation training, since its instructions are simplified and subjects may be more likely to continue to use the techniques after treatment. In the present study, we examined whether persons who suffer from recurrent anxiety can achieve improvement simply through a self-administered program that instructs them to develop their own method of getting relaxed at regular practice sessions.

Perhaps the major problem in using self-administered materials is the lack of client adherence to the program (Glasgow & Rosen, 1978). In Rosen, Glasgow, and Barrera's (1976) study of self-administered desensitization of snake phobias, less than half of the subjects in self-administered conditions completed half of their hierarchies, and the number of program steps completed was highly correlated with a variety of outcome measures. Thus, getting subjects to continue working on self-administered programs may increase their effectiveness.

Money deposits have been used extensively as a method of achieving compliance with particular aspects of treatment programs. Deposits have been found to be effective in getting clients to keep appointments (Grove & Fredricks, Note 1), lose weight (Hall, 1972), continue in treatment (Bellack, 1976; Bellack, Schwartz, & Rozensky, 1974; Ha-

gen, Foreyt, & Durham, 1976), and adhere to program procedures (Eyberg & Johnson, 1974). We could find no studies that evaluated the effects of a money deposit on persistence in a self-administered program. Therefore, we experimentally evaluated whether a money deposit would increase clients' practicing in the two self-administered relaxation programs that we wished to evaluate.

Finally, it would be useful to establish client characteristics that are associated with successful use of self-administered programs. One often-cited variable, locus of control (Rotter, 1966), has been suggested as a predictor of program adherence and outcome (Abramson, 1973; Balch & Ross, 1975; Friedman & Dies, 1974; Mahoney & Thoresen, 1974). In addition to the possibilities of prediction from locus of control, it may be that clients can predict their own success directly (cf. McReynolds & Stegman, 1976). A straightforward approach to the problem of predicting adherence, such as ratings of the likelihood that they will comply with instructions, may be the best and cheapest method of selecting those who will use a self-administered program successfully.

In summary, this study evaluates the effectiveness of two self-administered manuals for relaxation training as treatments for recurrent anxiety. The manuals are evaluated in comparison with a self-monitoring control condition. The usefulness of a money deposit in achieving program adherence and anxiety reduction is also evaluated. Finally, locus of control and clients' own predictions are evaluated as predictors of program adherence and outcome.

Method

Subjects

Subjects were recruited from the University of Oregon and a nearby community college. The project was described as offering treatment for anxiety and tension. At initial phone contact, 12 persons declined to participate, and 3 additional persons declined to participate at the first appointment. Among the remaining 61 subjects who entered the study, there were 38 females and 23 males. They ranged in age from 18 to 44 years, with a mean of 24 and a mode of 20 years.

Interventions

During the initial interview subjects were instructed on methods of self-monitoring their anxiety. Following a week of self-monitoring, subjects were randomly assigned to one of five conditions. Assignments were done separately for men and women in order to equate the proportion of men and women in each condition.

Self-administered progressive relaxation. Subjects in this condition received a manual on deep muscle relaxation, which was written by Rosen (1975, 1976). The manual contained complete instructions for learning to relax. In addition, it described methods of becoming able to relax in tension-arousing situations. The manual suggested that the client practice each day in two 15 to 20-minute sessions.

Self-administered progressive relaxation plus money deposit. Subjects in this condition received the Rosen manual and were asked to deposit an amount of money that they "wouldn't want to lose." The return of 25% of the deposit was contingent on subjects' returning their daily practice log, their self-monitoring records and a mail in questionnaire (see below). The remaining 75% of the deposit was returned contingent upon the subject's completing a minimum of 15 practice sessions. The manual recommends 42 sessions.

Client-directed relaxation. Subjects in this condition were given a two-page manual. The manual suggested that "practicing relating itself, regardless of the method . . . is the key element in feeling more relaxed and at ease." Subjects were asked to take the time to practice getting as relaxed as they could in two 15 to 20 minute sessions each day, using any method that they found effective. These subjects were also instructed to try to learn to become relaxed in the same way when anxiety-producing situations arise in real life.

Client-directed relaxation plus money deposit. Subjects in this condition received the manual just described and in addition were asked to put up a money deposit (identical to that for subjects in the self-administered progressive relaxation plus money deposit condition).

Self-monitoring. Subjects in this condition were asked to continue monitoring their daily anxiety experiences for an additional 2 weeks. In addition, they were reminded that there is evidence that self-monitoring helps some people change their behavior.

Instruments

Pre-treatment and post-treatment anxiety measures. During the initial interview all subjects completed the Spaulding State-Trait Anxiety Inventory and an assessment questionnaire that was devised for the purposes of this study. The State-Trait Anxiety Inventory (Spaulding, Gorsuch & Luthene, 1970) consists of two 20-item scales. The first is designed to measure respondents' anxiety at that moment,

and the second is designed to measure the respondents' tendency to be anxious in general. The pre-treatment assessment questionnaire asked subjects about previous experience with relaxation training, the importance of learning to relax, and the amount of therapist time that they would ideally like to have in learning to relax. They were then asked to list three situations that occurred at least once each week and during which they would like to be more relaxed. They rated their comfort in each of these situations at its most recent occurrence. Finally, the questionnaire asked them to rate the extent to which they believed that they could learn to relax using an entirely self-administered program. All ratings were done on 7-point scales.

The State-Trait Anxiety Inventory was administered at posttest, 3 weeks following the initial interview at the same time, and under the same circumstances as had existed at pretest. A postassessment questionnaire was also administered. It asked subjects to rate (a) their ability to relax, (b) their sense of self-control, (c) the benefit that they felt they had derived from the program, and (d) the likelihood that they would use the techniques that they had learned for relaxing in the future. Finally, subjects rated their comfort during the most recent occurrence of each of the three situations that they had specified at pretest.

Internal-external Locus of Control (I-E) Scale. A modified version of the I-E scale (Rotter, 1966) was administered at the initial session and at post-testing (Cohen, Rothbart, & Phillips, 1976). The scale was administered at posttest to determine whether outcome in the relaxation program was associated with changes in locus of control.

Self-monitoring. During the intake interview subjects were asked to identify two situations in which they experienced anxiety. One of the designated situations (Situation A) must have occurred daily. The second situation was one that occurred at least 3 times each week (Situation B). Clients were given pocket-size booklets and were instructed to record the occurrence of Situations A and B, their tension levels in these situations, and any other occurrences of tension. Ratings of tension were done on a 100-point scale, where 0 indicated no discomfort and 100 indicated the most extreme tension that the person had ever experienced. Subjects were also asked to rate their tension level for the entire day and to estimate the frequency of feeling tense during the day. All subjects continued self-monitoring for the entire 3 weeks of the program.

Expectancy questionnaire. Subjects who received relaxation instructions were asked to rate their expectations regarding the program immediately after they had read their relaxation instructions. They were asked to rate (a) how likely they thought the program would enable them to get relaxed, (b) how likely they thought they would do the prac-

¹ Copies of this manual are available from the second author.

sion sessions, and (c) how easy they thought it would be to follow the program. Ratings were made on 7-point scales.

Daily practice log. Subjects who received the relaxation instructions were asked to record the date, time, and length of each practice session, and were asked to rate on 10-point scales how easy it was to get relaxed and how relaxed they actually became.

Procedure

Subjects were seen on three occasions. During the first appointment the study was described and the subject was asked to read and sign a statement of informed consent. During this session, subjects filled out all pretreatment assessment instruments, were assisted in identifying situations that made them anxious (Situations A and B), and were given instructions on self-monitoring their anxiety. After 1 week of self-monitoring, subjects returned to the clinic and met with the experimenter. At this time subjects in the active treatment conditions were given relaxation manuals, and those in the deposit condition were asked to put up an amount of money that they "would not want to lose" (Grove & Fredricks, Note 1). Subjects in the relaxation conditions also received the expectancy questionnaire and practice logs. Those in the self-monitoring-only condition simply received further self-monitoring materials and were reminded that self-monitoring could help them to reduce anxiety. The third session occurred following 2 weeks of relaxation practice. At this time subjects were asked to complete the posttreatment instruments as described above. Subjects who had been in the self-monitoring-only condition received copies of the Rosen manual and instructions for its use at this session.

The amount of time that the experimenter spent with each subject was limited. The total time in which the experimenter was in contact with the subject over all three sessions was typically 1 hour and 40 minutes. The bulk of this time (approximately 60 minutes) was spent in subject's completing the relevant questionnaires. Thus, the procedures of this study approximate a wholly self-administered treatment program.

Results

Preliminary Analyses

Subjects' State Anxiety and Trait Anxiety scores were compared with published norms for these scales (Spielberger et al., 1970). On State Anxiety, this study's sample scored between the 73rd and 76th percentile for college undergraduates. For Trait Anxiety, the sample was at the 90th percentile. Comparisons with norms for clients seeking as-

sistance at a university counseling center indicated that the present sample was essentially equal to counseling center clients in State Anxiety and 6 points higher on Trait Anxiety. Of the 61 subjects, 39 had sought treatment previously. The mean length of time that clients had been anxious was 58 months.

As a check on randomization, one-way analyses of variance across the five treatments were conducted on all pretest measures. No significant differences were found (all $ps > .15$). One-way analyses of variance for sex differences at pretest and posttest were conducted as well as analyses of variance for the Sex \times Treatment Conditions interaction. For all but one measure, no sex-related differences were found (all $ps > .70$). The only significant sex-related difference was found for pretest and posttest locus of control scores (both $ps < .02$). Men scored significantly more internally than women, a difference that has been found by other investigators (Feather, 1968; Phares, 1976; Platt, Pomeranz, Eisenman, & DeLisser, 1970).

Statistical Analyses

Tests of the effects of the experimental manipulations were conducted by planned comparisons of change scores between pretesting and posttesting. These included (a) tests for differences between the two relaxation manuals; (b) tests for differences between the relaxation conditions and the self-monitoring-only condition; and (c) tests for differences between relaxation with money deposit and relaxation without a money deposit. These comparisons were made for both questionnaire anxiety measures and for self-monitored data. For self-monitored data, means were computed for Weeks 1 and 3, and change scores were derived by taking the differences between these means.

Effects of Relaxation Training

Table 1 presents basic pretreatment and posttreatment means for each condition.

Self-administered progressive relaxation versus client-devised relaxation. Planned com-

Table 1
Pretreatment and Posttreatment Means by Condition

Measure	Weeks 1 & 3	Progressive relaxation manual			Progressive relaxation manual & deposit			Client-devised relaxation			Client-devised relaxation & deposit			Self-monitoring- only control		
		n	M	SD	n	M	SD	n	M	SD	n	M	SD	n	M	SD
Self-monitoring																
Frequency of Situation A	Pre	9	8.89	7.41	6	8.67	5.57	9	8.56	3.01	9	9.44	5.22	9	7.33	5.61
	Post	9	8.78	7.66	6	3.83	3.87	9	5.00	4.42	9	6.78	3.96	9	4.33	2.50
Frequency of Situation B	Pre	9	7.00	5.02	6	11.17	10.53	9	8.67	7.40	9	3.78	2.44	9	5.44	6.73
	Post	9	5.78	5.81	6	4.17	2.86	9	3.11	2.80	9	3.67	3.50	9	6.56	7.97
M of Situation A	Pre	8	54.01	18.88	5	53.42	24.74	8	53.05	13.70	8	57.95	17.49	8	57.54	13.76
	Post	8	52.49	17.01	5	36.88	24.87	8	47.58	16.83	8	44.69	26.70	8	47.66	21.46
M of Situation B	Pre	7	54.17	20.59	4	62.72	25.99	7	48.41	15.83	6	54.27	10.84	6	59.06	15.44
	Post	7	54.70	13.07	4	37.33	24.09	7	47.41	16.99	6	35.03	23.66	6	52.98	23.04
Total frequency of all situations	Pre	9	32.00	9.31	6	33.33	22.21	9	34.22	20.67	9	21.56	7.16	9	33.67	35.78
	Post	9	29.89	8.43	6	24.83	15.97	9	19.78	12.61	9	20.00	8.96	9	39.56	42.97
Estimated frequency of all situations	Pre	9	49.78	29.10	7	74.43	86.32	9	54.11	53.78	9	55.11	85.32	9	47.89	43.44
	Post	9	41.11	18.31	7	30.71	13.70	9	34.00	26.63	9	35.78	30.90	9	61.78	72.65
M estimation of tension level	Pre	9	56.58	10.90	7	42.79	16.77	9	44.41	18.65	9	45.79	17.03	9	48.63	20.25
	Post	9	46.86	12.40	7	27.70	18.70	9	36.16	18.06	9	31.23	14.32	9	42.34	25.11
Pre-post measures																
STAI A-State	Pre	11	44.82	6.34	8	43.25	12.76	12	41.67	6.47	11	34.55	7.09	10	40.70	12.22
	Post	11	38.09	7.99	8	36.88	9.31	12	37.08	8.20	11	31.54	7.48	10	44.44	16.47
STAI A-Trait	Pre	11	52.64	5.95	8	54.75	6.21	12	48.25	7.84	11	48.36	8.09	10	50.40	9.35
	Post	11	42.82	9.58	8	44.88	12.10	12	42.25	9.13	11	41.09	8.19	10	45.30	11.02
Rating of Situation A	Pre	11	83.82	14.14	8	80.38	14.62	12	82.08	17.47	11	83.00	15.81	10	83.20	16.52
	Post	11	59.00	18.60	8	55.88	22.46	12	56.25	25.06	11	48.18	19.78	10	71.90	24.39
Rating of Situation B	Pre	11	83.27	13.61	8	82.88	11.84	12	79.50	15.76	11	82.73	8.48	10	79.10	20.98
	Post	11	57.72	18.22	8	45.00	23.76	12	52.75	25.65	11	50.46	19.55	10	64.30	32.12

Note. STAI = State-Trait Anxiety Inventory; A-Trait = Anxiety Trait scale; A-State = Anxiety State scale.

parisons of progressive and client-devised relaxation revealed no significant differences on any of the self-monitored or pre-post measures of anxiety. However, subjects who received progressive relaxation did rate themselves as able to relax at will to a significantly greater extent than did subjects in the client-devised condition, $t(47) = 2.26$, $p < .05$ (two-tailed). The groups did not differ on rated expectations.

Relaxation training versus self-monitoring only. Planned comparisons between relaxation training and self-monitoring-only conditions revealed a number of significant differences. Subjects who received relaxation training were significantly more improved on three self-monitored variables. They evidenced greater improvement on the frequency of Situation B, $t(13.9) = 2.25$, $p < .025$,² the total frequency of all anxiety-arousing situations, $t(14.6) = 2.46$, $p < .025$, and end of the day estimates of the total frequency of anxiety-arousing situations, $t(19.8) = 2.22$, $p < .025$. At pretreatment and posttreatment assessment, subjects receiving relaxation training were significantly more improved than the self-monitoring-only subjects on (a) State Anxiety scores, $t(47) = 2.40$, $p < .01$; (b) ratings of the discomfort that they experienced in Situation A, $t(47) = 1.78$, $p < .05$; and (c) ratings of their discomfort in Situation B, $t(47) = 1.78$, $p < .05$.

Three additional problem situations were rated at pretreatment and posttreatment but were not monitored during the program period. These unmonitored situations may be viewed as measures of the generalization effect produced by treatment conditions. Changes in the ratings of these situations did not differ among active treatment conditions. However, subjects in active treatment conditions evidenced significantly greater improvement on their ratings of discomfort in these situations than did the self-monitoring-only group, $t(47) = 1.73$, $p < .05$, $t(47) = 2.44$, $p < .05$, and $t(47) = 2.01$, $p < .05$.

Effects of Money Deposit

Subjects in the money deposit condition had to participate in a minimum of 15 practice sessions to get 75% of their money back.

Subjects in this condition were asked to deposit an amount of money that they "would not want to lose." Amounts deposited ranged from \$.10 to \$50. The mean amount of money deposited was \$6.01, and the median was \$1.50.

Practice and program adherence. Subjects in deposit and nondeposit conditions did not differ in the total number of practice sessions or in the average length of practice sessions. An examination of the group means for length of practice session indicated that subjects in the progressive relaxation without money deposit and the client-devised relaxation with deposit conditions spent significantly more time in practice sessions than did subjects in the other two groups, $t(31) = 2.60$, $p < .01$. Subjects in relaxation conditions rated their depth of relaxation in each practice session as well as the ease with which they were able to relax. Those in the money deposit condition had significantly higher ratings on ease, $t(32) = 2.60$, $p < .01$, and depth of relaxation, $t(32) = 2.54$, $p < .01$, than did nondeposit subjects.

Program adherence was assessed in terms of the number of days on which subjects kept self-monitored records, the number of steps in the program that they completed, and the number of days that they remained in the program. Deposit and nondeposit subjects did not differ on any of these variables. Nor were there significant differences for the proportion of subjects in each condition who dropped out of the study.

Changes in anxiety measures. Subjects in the money deposit condition improved significantly more than nondeposit subjects on two of the self-monitored ratings of anxiety. Deposit subjects improved significantly more on their ratings of tension in Situations A and B, $t(33) = 1.82$, $p < .05$; $t(5.1) = 2.63$, $p < .05$. There were no differences between deposit and nondeposit conditions on any of the questionnaire anxiety assessments.

Relation of amount deposited to other variables. Finally, it was found that the amount of money that subjects in the de-

² When the Bartlett-Box F was significant, separate variance estimates and degrees of freedom were calculated.

posit condition put up was correlated with their rating of the importance of their learning to relax, $r(25) = .40$, $p < .05$, but was not related to any outcome or practice measures.

Predictors of Program Adherence, Practice, and Outcome

Adherence and practice. Locus of control was significantly correlated with the number of days on which subjects self-monitored their anxiety, $r(61) = -.30$, $p < .025$, and the number of steps that they completed in the program, $r(61) = -.34$, $p < .025$. Locus of control was also significantly correlated with the number of days that the subject remained in the program, $r(61) = -.29$, $p < .025$. All of the correlations were in the direction of "internals" showing greater program adherence. Locus of control scores were not significantly related to practice variables such as number of relaxation sessions completed, length of practice sessions, or rated ease and depth of relaxation.

Subjects rated their expectations that they would (a) be able to learn to relax using the program, (b) do the practice sessions, and (c) find it easy to follow the program. These ratings were significantly correlated with both adherence and practice variables. With respect to adherence, subjects who predicted that the program would enable them to relax kept self-monitored records for more days, $r(43) = .38$, $p < .006$, and those who predicted that they would do the practice sessions kept self-monitored records for more days, $r(43) = .48$, $p < .001$, completed more steps of the program, $r(43) = .39$, $p < .005$, and remained in the program longer, $r(43) = .44$, $p < .002$. With respect to practice, it was found that subjects' ratings of the likelihood that they would practice were significantly correlated with the number of practice sessions that they completed, $r(43) = .45$, $p < .001$. Moreover, subjects who believed that the program would help them relax had significantly greater ease in relaxing, $r(36) = .38$, $p < .05$.

Outcome measures. The relationship of outcome to locus of control was evaluated by correlating locus of control scores with sub-

jects' change scores on the four pre-post anxiety measures and the seven self-monitor variables. Significant relationships were found for three variables. Those scoring in the internal direction on locus of control showed greater reduction in trait anxiety, $r(52) = -.28$, $p < .02$, and on self-monitored data they showed greater reductions in the total number of times that they were anxious per day, $r(42) = .27$, $p < .04$, and their end of day rating of tension, $r(43) = .30$, $p < .025$.

A similar analysis was conducted to determine whether subjects with high expectations for the program improved significantly more than subjects whose expectations were low. Subjects who predicted that the relaxation program would enable them to relax were significantly more improved on state, $r(40) = .50$, $p < .001$, and trait, $r(40) = .40$, $p < .006$, anxiety and on their daily rating of tension, $r(34) = .29$, $p < .05$. No other correlations were significant.

Program Adherence as a Predictor of Outcome

Finally, analyses were conducted to determine whether program adherence variables were related to outcome. No significant results were found. Differences among subjects in the number of steps that they completed in the program, the length and number of their relaxation practice sessions, their rated ease and depth of relaxation during these sessions, and the number of days on which they completed self-monitoring records were not related to any pre-post or self-monitored anxiety measures.

Discussion

The results of this study suggest that self-administered relaxation training, when combined with self-monitoring, can be of significant benefit to persons who suffer from recurring anxiety. Moreover, since few differences were found between progressive and client-devised relaxation groups, it appears that extensive, structured training in progressive relaxation is not necessary to achieve these benefits. Even though relaxation groups were superior to the self-monitoring-only condition, it should be noted that the latter

condition could itself be considered an active treatment. Indeed, clients in this condition improved on four of the seven self-monitored measures and on three of the four pre-post anxiety measures. Three of these improvements were statistically significant. This is consistent with other evidence in the literature that self-monitoring can produce changes in behavior (Nelson, Lipinski, & Black, 1975; Richards, McReynolds, Holt, & Sexton, 1976).

The improvements of the self-monitoring-only subjects also suggest that the beneficial effects of the relaxation conditions were dependent, in part, on these subjects' self-monitoring of anxiety. Relaxation training may not be of benefit in the absence of self-monitoring.

In both sets of relaxation training materials, subjects were explicitly instructed to attempt to actively use their relaxation when they began to feel anxious. Although it was not experimentally evaluated in the present study, this instruction may have been essential to the success of the program. The fact that relaxation subjects were significantly more improved than self-monitoring-only subjects in the unmonitored anxiety-arousing situations suggests that relaxation subjects were able to generalize the relaxation responses that they were learning to a variety of problematic situations.

It should be remembered that subjects in this study were college students. Although comparison with available norms suggests that they were moderately to highly anxious, the results of this study may not generalize to a more severe population such as chronically anxious outpatients in a clinical setting.

Although subjects in the relaxation training groups improved significantly more than those in the self-monitoring-only control condition, intersubject variability in improvement was quite high on many of the variables. Thus, even though the present findings point to the value of further developing relaxation treatments for recurrent anxiety, they do not represent clinically significant improvements for all relaxation subjects. One strategy for treating anxious clients would be to have a variety of techniques available.

The relaxation technique that has proven effective for the greatest number of people could be used first, but in the event that the first intervention was not successful, additional procedures should be available.

This study provides only limited support for the use of money deposits in self-administered programs. Despite the fact that the return of 75% of the money deposit was contingent on subjects' completing 15 or more relaxation practice sessions, money deposit groups did not practice significantly more than nondeposit groups. In fact, the mean number of practice sessions for deposit conditions was less than 15, which was lower than the comparable nondeposit conditions. Subjects who deposited money did rate the ease and depth of relaxation in practice sessions higher than did nondeposit subjects, but these results may be due to a nonsignificantly higher dropout rate among deposit subjects. Similarly, differential dropout may account for the few outcome differences that favored the nondeposit condition.

It would be premature, however, to completely abandon the use of money deposits in self-administered programs. Most subjects in the deposit condition put up very small amounts of money. The mean deposit was only \$6.01, and only six subjects put up 10 or more dollars. In a project to develop self-administered materials for depressed clients, we have generally found that persons seeking treatment put up larger amounts of money and generally do comply with our programs. Even though it is true that the amount of money deposited did not correlate with practice or outcome variables in this study, this result may be due to the fact that few subjects put up amounts of money that could have had an impact on their subsequent adherence to the program.

Our findings regarding the prediction of program success suggest the practical utility of simply asking people whether they think they will comply with and succeed in a self-administered program. The single rating of likelihood that subjects would complete the practice sessions was significantly related to the number of practice sessions completed as well as to all three measures of program ad-

herence. Similarly, the rating of the likelihood that the program would enable them to relax was significantly related to improvements on state anxiety and trait anxiety as well as daily tension ratings. Although locus of control scores were also related to program adherence measures and to three outcome measures, these relationships were less numerous and consistently less strong than those for the subjects' predictions about their own behavior. These results are consistent with the hypothesis that the I-E scale is itself measuring the extent to which people say that they will exert effort and can succeed on a variety of tasks. Thus, the scale need not be viewed as a measure of an internal, unitary construct. Rather it can be considered to assess the extent to which the people say, in a variety of ways, that their effort makes a difference. The scale's ability to predict success in self-administered programs may simply be due to the fact that working on a self-administered program is similar to the kinds of activities that the I-E scale asks about. Those who say that preparation makes the difference on a test will also say that they can use a self-administered program successfully. And, to some extent, these verbal responses are related to actual behavior (Skinner, 1957).

In concluding, we note several limitations of the present study that define some of the requirements for further research. First, the data were restricted to self-report questionnaires and self-monitoring. It is essential that we investigate whether such measures are accurately related to anxiety as it actually occurs in the daily life of clients. Second, the absence of relationship between program adherence and outcome variables leaves us unable to pinpoint the effective components of the relaxation treatment conditions. It would be useful to experimentally manipulate variables such as number of practice sessions and use of self-monitoring to elucidate their contribution to outcome. Finally, ratings of subject expectation for outcome were not obtained from self-monitoring-only subjects subsequent to their assignment to this condition. Since expectancy measures were related to outcome measures for subjects in relaxa-

tion conditions, it can be argued that the superiority of relaxation groups was more a matter of social influence than of the specific effects of relaxation and self-monitoring. It will be important in subsequent research to assess whether the self-monitoring condition does produce expectations for improvement that are comparable to those for relaxation conditions.

Reference Note

1. Grove, D. N., & Fredricks, H. D. *Teaching research parent training: A clinic model*. Unpublished manuscript, Teaching Research Infant and Child Center, Monmouth, Oregon, 1974.

References

- Abramson, E. E. A review of behavioral approaches to weight control. *Behaviour Research and Therapy*, 1973, 11, 547-556.
- Balch, P., & Ross, W. A. Predicting success in weight reduction as a function of locus of control: A unidimensional and multidimensional approach. *Journal of Consulting and Clinical Psychology*, 1975, 43, 119.
- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Bellack, A. S. A comparison of self-reinforcement and self-monitoring in a weight reduction program. *Behavior Therapy*, 1976, 7, 68-75.
- Bellack, A. S., Schwartz, J., & Rozensky, R. The contribution of external control to self-control in a weight reduction program. *Journal of Behavior Therapy and Experimental Psychiatry*, 1974, 5, 245-247.
- Benson, H., Beary, J., & Carol, M. The relaxation response. *Psychiatry*, 1974, 37, 37-45.
- Cohen, S., Rothbart, M., & Phillips, S. Locus of control and the generality of learned helplessness in humans. *Journal of Personality and Social Psychology*, 1976, 34, 1049-1056.
- Dawley, H. H. Anxiety reduction through self-administered relaxation. *Psychological Reports*, 1975, 36, 595-597.
- Eyberg, S. M., & Johnson, S. M. Multiple assessment of behavior modification with families: Effects of contingency contracting and order of treated problems. *Journal of Consulting and Clinical Psychology*, 1974, 42, 594-606.
- Feather, N. T. Change in confidence following success or failure as a predictor of subsequent performance. *Journal of Personality and Social Psychology*, 1968, 9, 38-46.
- Friedman, M. L., & Dies, R. R. Reactions of internal and external test-anxious students to counseling and behavior therapies. *Journal of Consulting and Clinical Psychology*, 1974, 42, 921.

- Gershman, L., & Stedman, J. M. Oriental defense exercises as reciprocal inhibitors of anxiety. *Journal of Behavior Therapy and Experimental Psychiatry*, 1971, 2, 117-119.
- Glasgow, R. E., & Rosen, G. M. Behavioral bibliography: A review of self-help behavior therapy manuals. *Psychological Bulletin*, 1978, 85, 1-23.
- Goldfried, M. R., & Trier, C. S. Effectiveness of relaxation as an active coping skill. *Journal of Abnormal Psychology*, 1974, 83, 348-355.
- Hagen, R. L., Foreyt, J. P., & Durham, T. W. The dropout problem: Reducing attrition in obesity research. *Behavior Therapy*, 1976, 7, 463-471.
- Hall, S. M. Self-control and therapist control in the behavioral treatment of overweight women. *Behaviour Research and Therapy*, 1972, 10, 59-68.
- Kass, D. J., Rogers, H. E., & Feldman, S. E. Deconditioning anxiety by individualized inhibitors. *Journal of Behavior Therapy and Experimental Psychiatry*, 1973, 4, 361-363.
- Kazdin, A. E., & Wilcoxon, L. A. Systematic desensitization and nonspecific treatment effects: A methodological evaluation. *Psychological Bulletin*, 1976, 83, 729-758.
- Mahoney, M. J., & Thoresen, C. E. *Self-control: Power to the person*. Monterey, Calif.: Brooks/Cole, 1974.
- McReynolds, W. T., & Stegman, R. Sayer vs. sign. *Behavior Therapy*, 1976, 1, 704-705.
- Nelson, R. O., Lipinski, D. P., & Black, J. L. The effects of expectancy on the reactivity of self-recording. *Behavior Therapy*, 1975, 6, 337-349.
- Paul, G. L. Inhibition of physiological response to stressful imagery by relaxation training and hypnotically suggested relaxation. *Behaviour Research and Therapy*, 1969, 7, 249-256.
- Phares, E. J. *Locus of control in personality*. Morristown, N.J.: General Learning Press, 1976.
- Platt, J. J., Pomeranz, D., Eisenman, R., & DeLisser, O. Importance of considering sex differences in relationships between locus of control and other personality variables. *Proceedings of the 78th Annual Convention of the American Psychological Association*, 1970, 5, 463-464. (Summary)
- Raskin, M., Johnson, G., & Rondestvedt, J. W. Chronic anxiety treated by feedback-induced muscle relaxation. *Archives of General Psychiatry*, 1973, 28, 263-267.
- Richards, C. S., McReynolds, W. T., Holt, S., & Sexton, T. Effects of information feedback and self-administered consequences on self-monitoring study behavior. *Journal of Counseling Psychology*, 1976, 23, 316-321.
- Rosen, G. M. A manual for self-administered relaxation. In J. P. Flanders (Ed.), *Practical psychology*. New York: Harper & Row, 1975.
- Rosen, G. M. *Don't be afraid: A self-instructional program for reducing fears and phobia*. New York: Spectrum Books, Prentice-Hall, 1976.
- Rosen, G. M., Glasgow, R. E., & Barrera, M. A controlled study to assess the clinical efficacy of totally self-administered systematic desensitization. *Journal of Consulting and Clinical Psychology*, 1976, 44, 208-217.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Sherman, A. R. Two-year follow-up of training in relaxation as a behavioral self-management skill. *Behavior Therapy*, 1975, 6, 419-420.
- Sherman, R., & Plummer, I. L. Training in relaxation as a behavioral self-management skill. *Behavior Therapy*, 1973, 4, 543-550.
- Skinner, B. F. *Verbal behavior*. New York: Appleton-Century-Crofts, 1957.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *Test manual for the State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Wolpin, M., & Raines, J. Visual imagery, expected roles and extinction as possible factors in reducing fear and avoidance behavior. *Behaviour Research and Therapy*, 1966, 4, 25-38.
- Zeisset, R. M. Desensitization and relaxation in the modification of psychiatric patients' interview behavior. *Journal of Abnormal Psychology*, 1968, 73, 18-24.

Received August 22, 1977 ■

Goal Definition by Staff Consensus: A Contribution to the Planning, Delivery, and Evaluation of Mental Health Services

Theodore W. Lorei and Eugene M. Caffey, Jr.
Veterans Administration, Washington, D.C.

To provide a basis for the evaluation of the Veterans Administration mental health services, a survey was conducted of staff opinion regarding the importance of several specific goals for these services. Nine goals were formulated and submitted to 6,435 central office and field facility staff to obtain their ratings of the importance of each of these goals, with an "of no importance" rating being possible. The goals dealing with (a) the development of patient skills necessary for being self-supporting, (b) the elimination of psychological disorders, and (c) the protection of patients and others from violence received the highest average ratings. The remaining goals were considered as having some importance, although there were substantial differences among the sample about the degree of importance. Interoccupational group differences in importance ratings were statistically significant but small. Although the goals formulated and ratified by staff were quite general, they were much more specific than previous goal statements. Because of this relative specificity, they provide useful guidance for planning, managing, delivering, and evaluating mental health services.

More than 6,000 professional staff members are involved in planning and delivering mental health services in Veterans Administration (VA) hospitals and clinics, as well as in the national administrative office. To what extent do these people have a common understanding of the objectives of these services? To find out, we surveyed their opinions about nine carefully formulated goals describing the desired impact of mental health services on the well-being of patients and communities. The results, analyzed for the group as a whole and for six professional subgroups, are reported here.

The authors acknowledge the substantial contribution of Lee Gurel to the formulation of the goal statements, as well as the assistance of Jack Collier, Robert Normand, Jacqueline McEwan, Wanda McKinney, the hospital and clinic survey coordinators, and of all the Veterans Administration staff who completed the questionnaire.

Requests for reprints should be sent to Theodore W. Lorei, Veterans Administration, 810 Vermont Avenue, N.W., Washington, D.C. 20420.

The study was developed from a previous application of decision theory concepts to decisions about releasing patients from psychiatric hospitals (Lorei, 1970). This application involved the formulation of a set of brief statements describing the possible outcomes of releasing or retaining patients. The importance of attaining or avoiding these outcomes was rated by clinical and administrative staff in 13 hospitals. These ratings suggested that staff differed substantially in their underlying objectives. For several reasons that are discussed more thoroughly later, it seemed important to inquire about these underlying objectives directly.

The research reported is related to the work just described but differs in several important respects. First, the central objects of the study were *goals* rather than *outcomes*. (For our purposes the relationship between outcomes and goals is one in which a goal is defined as a positively valued outcome.) Second, we are concerned not with a specific set

of alternative acts (release or retention from a hospital) but with a complex program of activities—all mental health services provided by the VA. And finally, in establishing the importance of goals for these services, staff from the entire VA mental health system, including those in the national administrative office, were consulted.

Although our original interest in outcomes and goals derived from decision theory, literature from several other areas also stresses the importance of goal clarity. Examples of these areas include management theory (Raia, 1974), education (Mager, 1975), psychotherapy (Mahrer, 1967), behavior modification (Bandura, 1969), and program evaluation (Weiss, 1972). The reason for the emphasis is clear: Individuals and groups cannot successfully manage organizations, devise curricula, modify the behavior of others, or evaluate programs unless the goals motivating the managers, the educators, the behavior modifiers, and the program operators are clear. The administration of mental health services (their planning, implementation, and evaluation) should profit from the clarification of goals found essential in so many fields.

The research described here was intended to contribute to the clarification and specification of the goals of all major mental health services. Our strategy was simple: We formulated a provisional set of goals and then asked the operators of the VA mental health services what they thought of these goals. The questions we sought to answer were (a) How important did VA staff consider each of the nine goals presented? (b) What additional goals did staff suggest? (c) How much consensus was there about the importance of each goal? and (d) What fundamental value dimensions appear to determine the goal importance ratings?

Method

Sample

The target sample included staff both in the VA national administrative office and in all hospitals and free-standing clinics serving substantial numbers of psychiatric patients. The national office sample included staff with responsibilities for mental health

programs. The hospital and clinic target sample was designed to include two subsamples: (a) management staff—hospital (or clinic) director and assistant director; the chief and assistant chief of staff; and the chiefs and assistant chiefs of psychiatry, psychology, social work and nursing and (b) clinical staff—all physicians (usually psychiatrists), psychologists, social workers, nurses, rehabilitation medicine senior therapists, and chaplains who spent at least 50% of their time working with psychiatric patients.

Instrumentation

Nine goals were written by the first author on the basis of the previous research on staff opinions about possible outcomes of release from or retention in psychiatric hospitals, review of agency documents, and discussions with VA staff in both the national office and in the field. The formulations were guided by Hatry's (1970) advice that program goals should emphasize outputs (in this case, *effects on people*) rather than inputs, such as number of patients treated, services provided, and so forth. For example, a statement such as "involve more patients and families in family therapy programs" was not regarded as a goal. However, included in the list of goal statements was one possible result of such programs—"to minimize stress on families resulting from living with and/or being responsible for veterans with psychiatric disabilities." The nine goal statements were written in questionnaire format. In the first section, the respondent was asked to rate the importance of each goal on a 6-point category scale. In the second section, goals were presented in sets of three, and the respondent was asked to rank them in order of their importance. Using a computer program developed by Gulliksen and Tucker (1961), which made the computation feasible, an incomplete pair-comparison procedure was used to derive interval scale values. (Only the results of the rating procedure are reported here.)

Results

Total Staff Opinions about Goal Importance

The mean importance ratings for all staff combined are shown in Table 1. Scale points ranged from 0 (of no importance) to 5 (extremely important). Three goals received an average rating greater than 4 (of great importance)—(a) To develop the self-care, interpersonal, and work skills necessary for veterans to become or remain self-supporting in the community; (b) to eliminate or reduce disorders of perception and thinking, severe emotional distress, drug addiction, and

Table 1
*Mean Importance Ratings for Goals of Services
 for Psychiatrically Disabled Veterans*

Goal	M	SD
How important is it for the VA to:		
4. Develop the self-care, interpersonal, and work skills necessary for veterans to become or remain self-supporting in the community?	4.62	.69
5. Eliminate or reduce disorders of perception and thinking, severe emotional distress, drug addiction, and alcoholism?	4.37	.88
6. Minimize the chances that severely disturbed veterans will injure or kill themselves or others?	4.21	1.09
9. Minimize restriction of veterans' personal liberty by requiring hospitalization only when absolutely necessary?	3.89	1.20
8. Provide a sheltered environment for those veterans unable to live in the community without serious distress or damage to themselves?	3.61	1.17
1. Teach skills useful for living within an institution (self-care, work details, etc.) to those patients who will probably always require institutional care?	3.57	1.21
2. Provide enough money to live in the community for those veterans who are too disabled to work?	3.27	1.28
3. Minimize stress on families resulting from living with and/or being responsible for veterans with psychiatric disabilities?	3.26	1.20
7. Minimize the chances that veterans will be "public nuisances" by virtue of odd behavior, chronic drunkenness, disorderly conduct, vagrancy?	2.72	1.42

Note. $N = 6,435$. VA = Veterans Administration. Scale points were defined as follows: 0 = of no importance; 1 = slightly important; 2 = moderately important; 3 = quite important; 4 = of great importance; 5 = extremely important.

alcoholism; and (c) to minimize the chances that severely disturbed veterans will injure or kill themselves or others. Even the lowest rated goal, "to minimize the chances that veterans will be 'public nuisances,'" was

rated 2.72 (between moderately and quite important).

Additional Goals

An informal content analysis was made of additional goals that 17% of the respondents had proposed (as they were invited to do on the questionnaire). Practically all suggestions described means for achieving the goals presented, such as expanding outpatient programs, supporting community living, providing greater continuity of treatment, and removing the requirement that only disabilities resulting from military service can be treated. The few suggested goals were not substantially different from the original set.

Staff Consensus

Total group. An indication of the agreement among staff about the importance of each goal is given by the standard deviations in Table 1. Theoretically, the standard deviations could range from 0 (perfect consensus) to 6 (equal numbers in the two extreme categories of the scale). Judged on this scale, the degree of staff agreement about all goals was quite high. The small deviations, of course, indicate greater consensus; and, as would be expected, there was more agreement about the high-rated goals than there was about the others.

Occupational subgroups. The deviations of each of the means of the goal importance ratings for six occupational groups from the means for the total staff group are presented in Table 2. This presentation makes it easy to see, for example, that of the six occupational groups, physicians regarded the development of skills necessary for self-support (Goal 4) as less important than did psychologists.

As can be seen from the F values (one-way analysis of variance) in Table 2, the mean importance ratings varied significantly across groups for all goals except "eliminate psychological disorders." Although most of the differences were small, there were medium or close to medium sized differences for the following three goals: (a) to minimize the chances that veterans will be public nuisances

Table 2
*Mean Goal Importance Ratings for All Staff and Deviations of Means
 for Occupational Subgroups*

Goal	Grand M	Chaplain	Nurse	MD	Psychol- ogist	RMST	SW	η^a	F
4. Develop skills necessary for self-support	4.62	-.11	.05	-.14	.10	.04	-.05	.12	17.64*
5. Eliminate psychological disorders	4.37	.07	.01	-.06	-.06	.08	.02	.05	2.78
6. Minimize chances of violence to self or others	4.21	.08	.16	.01	-.52	.03	-.13	.19	46.85*
9. Minimize restrictions on personal liberty	3.89	-.31	.20	-.33	.04	-.17	-.02	.17	38.51*
8. Provide a sheltered environment	3.61	.03	.09	-.12	-.36	.15	.05	.13	20.22*
1. Teach skills for institutional living	3.57	.15	.26	-.19	-.43	.27	-.35	.24	74.20*
2. Provide money	3.27	.16	.01	-.18	-.37	-.06	.40	.17	37.91*
3. Minimize family stress	3.26	.11	-.02	-.19	-.34	-.07	.49	.20	54.10*
7. Minimize chances that veterans will be "public nuisances"	2.72	.26	.23	.02	-.84	-.35	-.33	.25	84.44*

Note. $N = 6,292$. RMST = rehabilitation medicine senior therapists; SW = social workers. Scale points were defined as follows: 0 = of no importance; 1 = slightly important; 2 = moderately important; 3 = quite important; 4 = of great importance; 5 = extremely important.

^a Values of .10, .24, and .37 may be regarded as indicating small, medium, and large differences among occupational groups, respectively (Cohen, 1969).

* $p < .01$.

by virtue of odd behavior, chronic drunkenness, disorderly conduct, vagrancy; (b) to teach skills useful for living within an institution (self-care, work details, etc.) to those patients who will probably always require institutional care; and (c) to minimize stress on families resulting from living with or being responsible for veterans with psychiatric disabilities. Table 1 shows that these goals also had three of the highest standard deviations (measures of disagreement) for the total group.

Facilities. The mean goal importance ratings for the total staff group at each facility (hospitals and clinics) differed at a statistically significant ($p < .01$) level for all goals except "to eliminate psychological disorders," but, practically, these differences were small. There was most interfacility variation regarding the importance of minimizing restrictions on personal liberty and minimizing

chances that veterans will be public nuisances."

Value Determinants of Goal Importance Ratings

Although the nine goals were written to deal with relatively independent issues, it was apparent that there were interrelationships among them. To explore the possibility that some more general value orientations might underlie the way staff rated the importance of the nine goals, these goals were intercorrelated and factor analyzed. The varimax factor structure appears in Table 3.

Factor 1 is defined by a special concern with minimizing the chances of violent behavior and other behavior that would constitute a public nuisance. Factor 2 is characterized by attaching importance to minimizing family stress and providing money for

Table 3
Varimax Factor Structure

Goal	Factor				
	1	2	3	4	5
1. Teach skills for institutional living	.21	.11	.44	.11	.15
2. Provide money	.07	.48	.28	.08	.18
3. Minimize family stress	.16	.52	.08	.13	.09
4. Develop skills necessary for self-support	-.08	.06	.14	.47	.27
5. Eliminate psychological disorders	.25	.15	.03	.51	.04
6. Minimize chances of violence to self or others	.68	.09	.19	.11	.07
7. Minimize chances that veterans will be "public nuisances"	.54	.29	.32	.04	.00
8. Provide a sheltered environment	.31	.29	.50	.09	.07
9. Minimize restrictions on personal liberty	.05	.13	.09	.13	.42

Note. $N = 6,435$. This factor structure was obtained by rotating five principal factors extracted from a correlation matrix in which squared multiple correlations had been entered as communality estimates. The number of factors selected for rotation was suggested by examination of the latent roots.

disabled veterans to live in the community. Factor 3 emphasizes the importance of (a) providing a sheltered environment to patients, (b) teaching skills useful for institutional living to patients with poor prognoses, and (c) minimizing the chances that patients will be public nuisances. These goals have usually been termed *custodial*. Factor 4 is defined by the two goals dealing directly with patient change, one with the development of positive skills and the other with the elimination of psychopathology. These objectives can be considered as comprising what is usually considered a "treatment" ideology. Finally, Factor 5, defined by one goal, deals with the protection of patient liberties.

Discussion

It is unquestionably difficult to ascertain precisely the goals of people for either their personal or professional activities. As Quade (1975) has noted,

In stating objectives, officials do not always reveal what they really want, sometimes because they feel they must maintain a position or front, sometimes because it might mean loss of support, sometimes (maybe most often) because they don't know what they want (p. 86)

We suspect that this observation may have applied to some of our respondents. And even if they always had a clear conception of their goals (an unlikely supposition), it probably would be hard in many cases to express them on the brief instrument used. Nevertheless, having acknowledged these problems, we feel that our results provide at least a first approximation of the goals of VA staff for the mental health services that they provide.

The results suggest that a large proportion of staff considered the nine goals important, although their degree of importance varied considerably. Goals for producing patient change took precedence over those concerned with the welfare of relatives or others in the community and over those that implied acquiescence to permanent disability (such as teaching skills useful for institutional living). The most important objective involved making patients self-supporting, a position consistent with Jahoda's (1958) statement:

One value in American culture compatible with most approaches to a definition of positive mental health appears to be this: An individual should be able to stand on his own two feet without making undue demands or impositions on others. (p. xi)

The same value probably accounts for the

importance attached to teaching skills useful for institutional living. Even though to be institutionalized is to be dependent, there might be some opportunity for even an institutionalized person to "stand on his own two feet."

In addition to determining whether the goals that we formulated were considered important by staff, we were interested in the degree of staff consensus about goal importance. This interest derived from a perspective that has been well stated by Tawney (1926), writing in an entirely different context. He stated that

For the condition of effective action in a complex civilization is cooperation. And the condition of cooperation is agreement, both as to ends to which effort should be applied, and the criteria by which its success is to be judged. (p. 232)

The survey results suggest that agreement about the ends to which effort should be applied is fairly high. However, the generality of the goals may exaggerate the appearance of consensus. Some suggestion that this occurs is provided by studies by Prothro and Grigg (1960) and by McClosky (1964) that people's agreement with social issues is greater the more abstract the issue.

If the generality of the goal statement does mask important differences of opinion among staff, attention should be paid to even small indications of staff disagreement. For instance, the fact that psychologists rated seven of the nine goals as less important than did other occupational groups may point to important differences in values. Multidisciplinary discussion of these differences and of the special "preoccupations" of each profession—such as teaching institutional skills (nurses), developing independence (psychologists), and providing money (social workers)—might improve communication and strengthen commitment to common goals.

Although the nine goals are more specific than traditional "goals," such as providing treatment, rehabilitation, care, and custody, the process of specification needs to be continued. For example, the meaning of being "self-supporting" needs to be further clarified as to the prerequisite self-care, interpersonal, and work skills. Considerable guidance

could be obtained from educational literature on formulating instructional objectives. Mager (1975), for example, gives the following illustration: "Given a compass, ruler, and paper, [the student should] be able to construct and bisect any given angle larger than five degrees. Bisections must be accurate to one degree" (p. 79). This example includes a description of the behavior that is to occur as a result of instruction, the conditions under which it is to occur, and the standard of performance that is to be met. Such thinking seems to be directly translatable to the behavior that patients should manifest (or not) as a result of treatment.¹

Finally, it is probably worth noting that a commitment to clarifying the goals of mental health services does not imply preference for any particular paradigm for conceptualizing abnormal behavior, for example, medical model, psychoanalytic, or behavioristic. The focus of concern about goal clarity is on the end product of treatment and not the intermediate conditions (such as development of insight) for achieving it. Perhaps the greatest contribution of the goal-oriented approach is that it provides a common framework for communications among people of diverse professional and theoretical positions.

¹ Although space limitations prevent describing our applications of the results of the goal-setting process reported here to the evaluation of day hospitals, day treatment centers, and drug-dependence treatment centers, further information can be obtained from the first author.

References

- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Cohen, J. *Statistical power analysis for the behavioral sciences*. New York: Academic Press, 1969.
- Gulliksen, H., & Tucker, L. R. A general procedure for obtaining paired comparisons from multiple rank orders. *Psychometrika*, 1961, 26, 173-183.
- Hatry, H. P. Measuring the effectiveness of non-defense public programs. *Operations Research*, 1970, 18, 772-784.
- Jahoda, M. *Current concepts of positive mental health*. New York: Basic Books, 1958.
- Lorei, T. Staff ratings of the relative importance of the consequences of release from or retention in

- a psychiatric hospital. *Journal of Consulting and Clinical Psychology*, 1970, 34, 48-55.
- Mager, R. F. *Preparing instructional objectives*. Belmont, Calif.: Fearon, 1975.
- Mahrer, R. R. *The goals of psychotherapy*. New York: Appleton-Century-Crofts, 1967.
- McClosky, H. Consensus and ideology in American politics. *American Political Science Review*, 1964, 58, 361-362.
- Prothro, J., & Grigg, C. Fundamental principles of democracy: Bases of agreement and disagreement. *Journal of Politics*, 1960, 22, 276-294.
- Quade, E. S. *Analysis for public decisions*. New York: American Elsevier, 1975.
- Raia, A. P. *Managing by objectives*. Glenview, Ill.: Scott, Foresman, 1974.
- Tawney, R. H. *Religion and the rise of capitalism*. New York: New American Library, 1926.
- Weiss, C. *Evaluation research: methods of assessing program effectiveness*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Received October 11, 1977 ■

Comparison of Electromyographic Feedback and Progressive Relaxation Training in Treating Circumscribed Anxiety Stress Reactions

Martha P. Miller
University of Oklahoma

Philip J. Murphy
Oklahoma State University

Terry P. Miller
Stillwater, Oklahoma

This study examined the effects of electromyographic (EMG) feedback and progressive relaxation training on the anxiety stress reactions of patients having recurrent, negative reactions to dental treatment. Twenty-one subjects selected from the patient files of one dentist were randomly assigned to one of three groups: EMG feedback, progressive relaxation, or control. Four dependent measures, EMG level, Dental Anxiety Scale (DAS), and State-Trait Anxiety Inventory (A-State and A-Trait), were collected at prerelaxation and postrelaxation training dental appointments. Results showed significant, comparable decreases in EMG levels across dental appointments for both EMG feedback and progressive relaxation groups but not for the control group. On the DAS and A-State measures, significant decreases in all groups were found. Although the decreases shown by the EMG feedback and progressive relaxation groups did not differ significantly from each other, they were both significantly greater than the decrease shown by the control group.

Even though human reaction to stress has been the focus of much psychological investigation, there is little research in the area that has been carried out under natural, stress-provoking conditions. Lazarus (1966) has noted that the dental context provides an excellent area in which to study both physiological and psychological aspects of stress. Due to the nature of the procedures involved, anxiety stress reactions are experienced by many patients undergoing dental treatment. The dental setting thus provides the opportunity to study firsthand the effectiveness of psychotherapeutic techniques aimed at relieving stress reactions.

Relaxation of the body musculature has

been suggested as one means by which anxiety stress reactions can be reduced (Jacobson, 1938; Shultz & Luthe, 1959; Wolpe, 1958). Emphasizing the important role of relaxation in treating a number of stress-related disorders including anxiety states, Jacobson (1938) developed a verbal technique (progressive relaxation) designed to systematically train individuals to relax. More recently, relaxation training using electromyographic (EMG) feedback has been suggested as a way to reduce anxiety stress reactions (Budzynski & Stoyva, 1969; Green, Green, & Walters, 1973).

Several studies have dealt with whether muscular relaxation learned through progressive relaxation training leads to the reduction of anxiety. Comparing the effects of progressive relaxation and hypnosis, Paul (1969) found that anxiety indicators in normal subjects were reduced significantly in the relaxation group in contrast to the hypnosis and control groups. Partial substantiation of the relaxation training-anxiety reduction

This article is based on a dissertation submitted by the first author to the University of Oklahoma in partial fulfillment of the requirements for the PhD degree.

Requests for reprints should be sent to Martha P. Miller, who is currently in the private practice of psychology at 4900 North Portland, Suite 112, Oklahoma City, Oklahoma 73112.

hypothesis was shown in studies by Wilson and Wilson (1970) and Connor (1974). In the former study, only the high-anxiety relaxation group showed significant decreases in anxiety indicators, whereas in the latter study, the relaxation group showed reductions in physiological (but not self-report) stress indicators. Although progressive relaxation has long been held to be an effective treatment for reduction of anxiety stress reactions (Haugen, Dixon, & Dickel, 1958; Jacobson, 1938, 1970), little research has been directed toward the empirical validation of this technique as an effective treatment mode.

The assessment of EMG feedback as a treatment for the reduction of anxiety stress reactions has been the focus of several recent investigations. A single group study by Raskin, Johnson, and Rondetvedt (1973) indicated that EMG feedback is of limited value in the treatment of chronic generalized anxiety. Stoyva and Budzynski (1974), however, reported successful use of EMG feedback in the clinical treatment of "several dozen" pervasive anxiety patients. Several controlled studies comparing the effectiveness of EMG feedback and modified types of progressive relaxation training were in general agreement that EMG feedback was superior to progressive relaxation training with regard to speed of learning and depth of relaxation obtained (Coursey, 1975; Haynes, Moseley, & McGowan, 1975; Reinking & Kohl, 1975). These latter studies contained normal subjects who were not involved in a stress situation. It is clear that the comparative effectiveness of EMG feedback and progressive relaxation training in reducing anxiety under natural, stress-provoking circumstances has yet to be determined.

The purpose of this study was twofold—(a) to ascertain the effects of relaxation training on anxiety stress reactions in a dental setting and (b) to determine the comparative effectiveness of EMG feedback and progressive relaxation in relieving anxiety stress reactions of a relatively circumscribed, though reoccurring, nature. The following hypotheses were tested:

1. EMG feedback and progressive relaxation training will lead to significant decreases

in anxiety stress reactions as compared to self-relaxation control procedures.

2. EMG feedback relaxation training will produce significantly greater stress reduction than will progressive relaxation training.

Method

Subjects

A list of patients classified as prone to anxiety stress reactions in a dental setting was compiled by a dentist from his total active file of patients on the basis of the dentist's observations over time and also on patients' verbal reports to him. The patients were contacted by phone by the dentist to determine their interest in participating in a study of treatment techniques aimed at the reduction of stress reactions in dental situations. Those patients who were interested were scheduled for an initial interview that included a brief medical-dental history and a dental examination. Any patient who either was currently using drugs that might effect the results of the study or who was being seen regularly by other health service providers was excluded from the study. The subjects who were included required dental work of a simple restorative nature entailing at least two separate dental appointments. The final list included 21 subjects (17 females, 4 males). The disproportionate male/female ratio resulted from two factors: (a) Twice as many females ($n = 28$) as males ($n = 14$) were originally identified by the dentist as stress prone in a dental setting, and (b) the males who were contacted were more reluctant to participate in the study than were the females. Subjects ranged in age from 21 to 48 years, with a mean age of 35 years.

Apparatus

EMG measures were recorded with an Autogen 1500 feedback myograph using standard frontal placements 2 inches (5.08 cm) on either side of the center of the forehead and 1 inch (2.54 cm) above each eyebrow (Venables & Martin, 1967). A ground electrode was secured to the forehead midway between the other electrodes.

Connected to the Autogen unit were stereophonic headphones through which subjects in the EMG feedback group received auditory feedback of ongoing muscular tension. This feedback was presented in the form of clicks that were logarithmically proportional to the level of EMG activity being monitored. All meter readings were based on average integral microvolts.

Procedure and Measures

During the first phase of the initial dental appointment in which actual dental work occurred,

baseline frontalis EMG readings were taken. After a 3-min adaptation period while the patient was seated in the dental chair, the baseline EMG readings were recorded once every 10 sec for a 3-min period and were averaged to obtain a single score. After removal of the electrodes, each subject completed a set of self-report measures including the Dental Anxiety Scale (DAS; Corah, 1969) and the State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1970). Upon completion of these measures, the scheduled dental treatment began. At the conclusion of the initial dental appointment, all subjects were scheduled for 10 training or control sessions to extend over a 4-week period. Both dental appointments and all training sessions were scheduled at approximately the same time of day, with a maximum deviation of 2 hours for any 1 subject. Subjects were then randomly assigned to one of three groups ($n=7$ each): EMG feedback training, progressive relaxation training, or self-relaxation control.

Subjects in the progressive relaxation group received 10 training sessions 10–40 minutes in length in the manner standardized by Bernstein and Borkovec (1973). The longer time periods were necessary for completion of training during the first three sessions, with session length decreasing as the training progressed. All training was completed while the subjects' eyes were closed. The progressive relaxation training was "live," rather than prerecorded on tape, as Paul and Trimble (1970) have shown the former method to be significantly more effective in reducing stress responses.

In the EMG feedback group, subjects also received 10 training sessions. Each session was 20 minutes long (Budzynski & Stoyva, 1969). EMG subjects were told that they would be hearing clicks through the headphones that would be proportional to their moment-to-moment muscular tension (i.e., the clicks would increase in speed as their tension level increased and would decrease in speed as they relaxed). EMG subjects were then instructed to close their eyes and to slow down the speed of the clicks.

As with subjects in the treatment groups, control subjects met with the experimenter for 10 sessions. Each session was 20 minutes in length. The control subjects were told, as were subjects in the other groups, that they were participating in a study designed to see whether practice in relaxation might help people feel more comfortable during future dental appointments. After being instructed to close their eyes, control subjects were asked to relax themselves as best they could. These subjects were not informed of the control aspect of their participation.

Treatment and control sessions were conducted in the physical therapy room of a physician's office adjacent to the office of the participating dentist. The room was well insulated from sound on all sides. During relaxation sessions all subjects reclined on a hospital bed with the experimenter seated in a nearby chair. In all cases, the second dental appointment was scheduled within 2 weeks after completion of the 10 relaxation sessions. EMG recordings of

muscle tension levels and self-report data were collected in an identical manner as that collected during the first dental appointment. The dentist who recorded EMG levels and collected self-report data was unaware of the experimental group to which each patient was assigned.

Routine dental checkup notices were mailed to all subjects approximately 1 year after their second experimental dental appointment. Those subjects who responded to this call for checkup appointments and who on examination required dental treatment other than routine cleaning were scheduled for a follow-up appointment. EMG levels and self-report data were collected in the same manner as in the first and second dental appointments.

Results

EMG Levels

Training sessions. The mean EMG levels in microvolts for all groups were determined for each of the 10 training or control sessions that occurred during the period between dental appointments. Figure 1 shows the EMG trends for each group across these 10 sessions. If it is assumed that learning to relax is accompanied by progressively lower EMG levels across sessions, significant linear trends for each group would be one indicator that learning had occurred. Separate tests for trends across the 10 training sessions for each group showed significant linear trends for the progressive relaxation, $F(1, 18) = 8.93, p < .01$, and the EMG feedback, $F(1, 18) = 6.31, p < .05$, groups but not for the control group (F

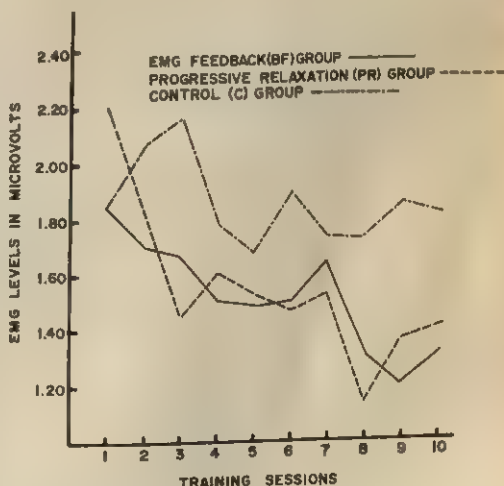


Figure 1. Trends for training session electromyogram (EMG) levels.

Table 1

Analysis of Variance of the Effects of EMG Feedback, Progressive Relaxation, and Self-Relaxation on Training Session EMG Levels with Planned Trend Comparisons

Source	df	MS	F
Groups (A)	2	2.07	.52
Error (within groups)	18	3.98	
Training sessions (B)	9	.63	5.73**
Linear trend	1	4.54	34.92***
A × B	18	.19	1.73 *
Differences in linear trends	2	.30	2.31
Error (B × Subjects within Groups)	162	.11	
Linear error	18	.13	

Note. EMG = electromyograph.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

= .46). Further analysis of the data was accomplished through use of a 3×10 (Groups × Training Sessions) analysis of variance (ANOVA), with the results being presented in Table 1. Of particular interest is the Groups × Training Session interaction, which indicates significant differences in trends among groups. Although a significant difference among linear trends would be an expedient way of showing differential learning among groups, it can be seen from Table 1 that this linear trend difference was not significant. As a result, trends other than linear must be considered in order to account for the significant Groups × Training Sessions interaction. Therefore, although significant differences among linear trends were not shown, the progressive relaxation and EMG feed-

back groups did show significant linear decreases in EMG levels while the control group did not.

Dental appointments. The mean EMG level in microvolts for each group across dental appointments is presented in Table 2, as are the results of a 3×2 (Groups × Appointments) ANOVA. A significant decrease in EMG level across dental appointments was found. Even though the Groups × Appointments interaction was not significant, planned analysis of the data using an ANOVA for simple main effects showed that significant decreases in EMG levels from the first to the second dental appointment occurred in the EMG feedback ($p < .05$) and the progressive relaxation ($p < .05$) groups but not in the control group. Although EMG levels for both the

Table 2

Analysis of Variance of the Effects of EMG Feedback, Progressive Relaxation, and Self-Relaxation on Dental Appointment EMG Levels

Source	df	MS	F	Condi- tion	Appointment <i>M</i>	
					1	2
Groups (A)	2	.71	.23	BF	3.50	2.08
Error (within groups)	18	2.56				
Appointments (B)	1	12.27	8.13*	PR	3.70	1.86
A × B	2	3.58	2.37			
Error (A × Subjects within Groups)	18	1.51		C	2.35	2.43

Note. EMG = electromyogram; BF = EMG feedback; PR = progressive relaxation; C = control.

* $p < .05$.

Table 3
Analysis of Variance of the Effects of EMG Feedback, Progressive Relaxation, and Self-Relaxation on Dental Anxiety Scale Scores

Source	df	MS	F	Condition	Appointment M	
					1	2
Groups (A)	2	23.74	3.05	BF	14.00	8.00
Error (within groups)	18	7.79				
Appointments (B)	1	201.53	70.20**	PR	12.43	7.71
A × B	2	11.45	3.99*			
Error (A × Subjects within Groups)	18	2.87		C	13.86	11.43

Note. EMG = electromyogram; BF = EMG feedback; PR = progressive relaxation; C = control.

* $p < .05$.

** $p < .001$.

EMG feedback and progressive relaxation groups were reduced significantly across dental appointments, these decreases were not significantly different from each other.

Dental Anxiety Scale

Table 3 consists of the mean score on the DAS for each group across appointments and the results of a 3×2 (Groups × Appointments) ANOVA. A significant decrease in DAS scores occurred between dental appointments, as demonstrated by the significant main effect. Further, a significant Groups × Dental Appointments interaction was also shown. A planned ANOVA for simple main effects accounted for the interaction effect by demonstrating that first, all groups evidenced a significant decrease in scores across appointments (for EMG feedback, $p < .001$; for progressive relaxation, $p < .001$; for controls, $p < .05$), and, second, the groups' scores, although not differing on the first dental appointment, did differ significantly on the second ($p < .05$). Individual comparisons indicated that even though there was no significant difference between the EMG feedback and progressive relaxation scores at the second dental appointment, the mean of these scores differed significantly ($p < .05$) from that of the control group. A t test on the difference scores between appointments for the EMG feedback and progressive relaxation groups was not significant. Thus, DAS scores for all groups showed significant reductions; however, the reductions shown by the EMG

feedback and progressive relaxation groups, while not differing significantly from each other, were significantly lower than those in the control group.

STAI-State. In Table 4 the mean STAI A-State scores for each group and the results obtained from a 3×2 (Groups × Appointments) ANOVA are presented. Main effects for groups ($p < .05$) and appointments ($p < .001$) and the interaction effect ($p < .001$) were significant. A planned ANOVA for simple main effects revealed significance for groups at both levels of appointments and for appointments at all levels of groups. Individual comparisons showed that at the first appointment, the score of the progressive relaxation group was significantly lower than that of the EMG biofeedback group ($p < .05$); however, the score of neither the progressive relaxation group nor the EMG biofeedback group differed significantly from that of the control group. At the second dental appointment, the score of the EMG biofeedback group did not differ significantly from that of the progressive relaxation group, although the scores of both the EMG biofeedback ($p < .01$) and the progressive relaxation groups ($p < .01$) differed significantly from that of the control group. A t test of difference scores between appointments for the EMG biofeedback and progressive relaxation groups was not significant. Thus, all groups showed significant decreases in A-State scores; however, even though the reductions shown by the EMG biofeedback and progressive relaxation groups did not differ significantly from each

Table 4

Analysis of Variance of the Effects of EMG Feedback, Progressive Relaxation, and Self-Relaxation on A-State Scores

Source	df	MS	F	Condition	Appointment <i>M</i>	
					1	2
Groups (A)	2	848.86	5.77*	BF	55.43	30.57
Error (within groups)	18	147.14				
Appointments (B)	1	2,288.09	2,138.40**	PR	43.43	27.43
A × B	2	405.82	379.27**			
Error (A × Subjects within Groups)	18	1.07		C	52.71	49.29

Note. EMG = electromyogram; A-State = State Anxiety scale of the State-Trait Anxiety Inventory; BF = electromyogram feedback; PR = progressive relaxation; C = control.

* $p < .05$.

** $p < .001$.

other, they were both significantly greater than the decreases shown by the control group.

Trait Anxiety. Table 5 includes mean Trait Anxiety scores for all groups across dental appointments and the results of a 3×2 (Groups × Appointments) ANOVA. A significant reduction was shown between scores on the first and second dental appointments. However, the Groups × Appointments interaction was not significant. Even though the subjects reduced their trait anxiety from the first to the second dental appointments, there was no differentiation among the three groups in the trait anxiety reduction effect.

Discussion

Both the training session data and the pre-post appointment data lend support to the

first hypothesis, which contends that EMG feedback and progressive relaxation training will lead to significant reductions in anxiety stress reactions as compared to self-relaxation control procedures. The second hypothesis, which holds that EMG feedback relaxation training will produce significantly greater anxiety stress reduction than progressive relaxation training, was not confirmed. Both treatment groups were equally effective in reducing transitory, situational anxiety such as that found in the dental setting.

A review of the data suggests that patients having recurrent anxiety stress reactions in dental settings may obtain a significant degree of relief from relaxation training with either EMG feedback or progressive relaxation techniques. The EMG levels of both the feedback and the progressive relaxation

Table 5

Analysis of Variance of the Effects of EMG Feedback, Progressive Relaxation, and Self-Relaxation on A-Trait Scores

Source	df	MS	F	Condition	Appointment <i>M</i>	
					1	2
Groups (A)	2	88.10	.54	BF	35.71	31.00
Error (within groups)	18	159.88				
Appointments (B)	1	77.36	5.49*	PR	36.29	34.71
A × B	2	10.57	.75			
Error (A × Subjects within Groups)	18	14.08		C	39.29	37.43

Note. EMG = electromyogram; A-Trait = Trait Anxiety scale of the State-Trait Anxiety Inventory; BF = EMG feedback; PR = progressive relaxation; C = control.

* $p < .05$.

groups showed significant decreasing linear trends across training sessions suggestive of learning. No such trend, however, was observed in the control group. Also, from the first to the second dental appointments, the feedback and the progressive relaxation groups showed significant decreases in scores on three of four dependent measures (EMG, DAS, A-State). Even though significant decreases were also shown by the control group on the latter two measures, these decreases were significantly less than those shown by the EMG feedback and progressive relaxation groups. No indication of a clear-cut superiority of one type of relaxation training over the other was shown on the EMG, DAS, or A-State measures as pertaining to situational anxiety. Rather, on the basis of this evidence, it would seem that patients in both the EMG feedback and progressive relaxation groups were more comfortable, physiologically and psychologically, when exposed to the immediate threat of dental stimuli after receiving relaxation training.

The consistency of findings for both treatment groups provides the suggestion of the mechanism for the reduction effect of the anxiety of the treatment groups. The learned reduction in EMG levels across sessions, shown by the decreasing linear trends, transferred to the posttreatment dental appointment. In addition to the reduced EMG levels that the patients exhibited in the dental chair immediately prior to the dental work, they reported less dental anxiety and state anxiety at that time. Apparently, the learned reduction in muscular tension that was a result of either the EMG feedback or the progressive relaxation procedures produced the reductions in both physiological and self-report measures of anxiety.

An important point to note is that the research under discussion here is a clinical treatment study. This represents a major difference between it and other studies that have investigated the comparative effects of relaxation techniques, notably the Coursey (1975) study and the Reinking and Kohl (1975) study, both of which contained normal subjects. As mentioned earlier, the subjects involved in the current study were tested under natural conditions, which to them were

highly stressful. Beyond the observations of the dentist who identified the dental-anxious subjects, representing less than 1% (original subject pool $N = 42$) of an available patient population ($N > 5,000$), other factors pointed to a high discomfort level among the subjects selected. Indicative of this discomfort were pretraining DAS scores and A-State scores, which ranked at about the 90th percentile when compared to scores found in normal populations (Corah, 1969; Spielberger et al., 1970). Further, the pretraining muscular tension levels of the subjects were approximately 32% higher than the pretraining levels of the normal subjects in the Coursey (1975) study and 23% higher than those in the Reinking and Kohl (1975) study. (For the purpose of comparison, the peak-to-peak microvolt readings reported in the latter two studies were divided by the constant 3 in order to approximate the average integral microvolt readings used here.)

In conclusion, direct clinical implications that lead from the results of this study point to EMG feedback relaxation training and live, therapist-directed progressive relaxation training as effective treatments for relatively short-lived but recurrent bouts of anxiety that are bound to a particular stimulus event.

References

- Bernstein, D. A., & Borkovec, T. D. *Progressive relaxation training: A manual for the helping professions*. Champaign, Ill.: Research Press, 1973.
- Budzynski, T. H., & Stoyva, J. M. An instrument for producing deep muscle relaxation by means of analog information feedback. *Journal of Applied Behavior Analysis*, 1969, 2, 231-237.
- Connor, W. H. Effects of brief relaxation training on automatic response to anxiety-evoking stimuli. *Psychophysiology*, 1974, 11, 591-599.
- Corah, N. L. Development of a dental anxiety scale. *Journal of Dental Research*, 1969, 48, 596.
- Coursey, R. D. Electromyograph feedback as a relaxation technique. *Journal of Consulting and Clinical Psychology*, 1975, 43, 825-834.
- Green, E. E., Green, A. M., & Walters, E. D. Biofeedback training for anxiety tension reduction. *Annals of New York Academy of Science*, 1973, 233, 157-161.
- Haugen, G. B., Dixon, H. H., & Dickel, H. A. *A therapy for anxiety tension reactions*. New York: MacMillan, 1958.
- Haynes, S. N., Moseley, D., & McGowan, W. T. Relaxation training and biofeedback in the re-

- duction of frontalis muscle tension. *Psychophysiology*, 1975, 12, 547-551.
- Jacobson, E. *Progressive relaxation*. Chicago: University of Chicago Press, 1938.
- Jacobson, E. *Modern treatment of tense patients*. Springfield, Ill.: Charles C Thomas, 1970.
- Lazarus, R. S. Some principles of psychological stress and their relation to dentistry. *Journal of Dental Research*, 1966, 45, 1620-1626.
- Paul, G. L. Physiological effects of relaxation training and hypnotic suggestion. *Journal of Abnormal Psychology*, 1969, 74, 425-437.
- Paul, G. L., & Trimble, R. W. Recorded vs. "live" relaxation training and hypnotic suggestion: Comparative effectiveness for reducing physiological arousal and inhibiting stress response. *Behavior Therapy*, 1970, 1, 285-302.
- Raskin, M., Johnson, G., & Rondestvedt, J. W. Chronic anxiety treated by feedback-induced muscle relaxation: A pilot study. *Archives of General Psychiatry*, 1973, 28, 263-266.
- Reinking, R. H., & Kohl, M. L. Effects of various forms of relaxation training on physiological and self-report measures of relaxation. *Journal of Consulting and Clinical Psychology*, 1975, 43, 595-600.
- Shultz, J., & Luthe, W. *Autogenic training: A psychophysiological approach to psychotherapy*. New York: Grune & Stratton, 1959.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. *Manual for the State-Trait Anxiety Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1970.
- Stoyva, J., & Budzynski, T. Cultivated low arousal—An antistress response? In L. V. DiCara (Ed.), *Limbic and autonomic system research*. New York: Plenum, 1974.
- Venables, P. H., & Martin, I. (Eds.). *A manual of psychophysiological methods*. Amsterdam: North Holland, 1967.
- Wilson, A., & Wilson, A. S. Psychophysiological and learning correlates of anxiety and induced muscle relaxation. *Psychophysiology*, 1970, 6, 740-748.
- Wolpe, J. *Psychotherapy by reciprocal inhibition*. Stanford, Calif.: Stanford University Press, 1958.

Received October 18, 1977 ■

Hemispheric Asymmetry and Jewish Intelligence Test Patterns

Colin Martindale
University of Maine

On the basis of intelligence test performance, Dershowitz and Frankel have hypothesized that Jewish subjects are less psychologically differentiated (more field dependent) than Protestant and Catholic subjects. It is argued that differential intelligence test patterns are better explained in terms of differential emphasis on abilities mediated by the left and right cerebral hemispheres. The hypothesis is advanced that Jewish subjects exhibit a tendency toward left-hemisphere dominance in comparison to Protestant and Catholic subjects. Evidence for, and implications of, the hypothesis are discussed.

Dershowitz and Frankel (1975) have recently summarized the results of a number of studies showing that Jewish subjects tend to be characterized by relatively low scores on some of the performance subtests of the Wechsler Intelligence Scale for Children (WISC) and of the Wechsler Adult Intelligence Scale. Scores on Picture Completion, Picture Arrangement, Block Design, and Object Assembly are low in relation to scores on the verbal subtests of Comprehension, Information, Arithmetic, and Similarities. Studies are cited as showing a similar but less extreme pattern in Protestant children and an even less extreme pattern in Catholic (Irish and Italian) children. Dershowitz and Frankel related these patterns to Witkin's (1967) concept of psychological differentiation. According to their hypothesis, Jewish subjects are less psychologically differentiated than Protestant or Catholic subjects.

Even though low levels of psychological differentiation have been shown to be related to poor performance on three of the subtests on which Jewish subjects exhibit relative deficits, Dershowitz and Frankel (1975) themselves admit that they are unable to account for the poor Picture Arrangement scores of Jewish subjects with this explanation. Further, they cite only one unpublished study (Litman,

Note 1)—in which a version of the Embedded Figures Test was used—that would directly support the differentiation hypothesis. Two published studies (Dershowitz, 1971; Wendt & Burwell, 1964) reported failures to find significant differences between Jewish and non-Jewish children on Embedded Figures Test performance, although differences were in the expected direction.

Dershowitz and Frankel's (1975) findings are more parsimoniously explained by the hypothesis that Jewish subjects show a relative superiority on tasks that are dependent on the left cerebral hemisphere and a relative deficit on tasks that are dependent on the right cerebral hemisphere, whereas Protestants show a similar but less extreme pattern and Catholics are relatively balanced in their performance. Recent research using a number of strategies—such as split visual-field stimulation, dichotic listening, and recording electroencephalogram during task performance—has led to the hypothesis that the left hemisphere is specialized for linguistic, analytical, and sequential tasks, whereas the right hemisphere is specialized for tasks requiring holistic, spatial abilities. To my knowledge, no one has used these strategies to test for differential performance as a function of subjects' religious backgrounds. However, a number of studies of the intellectual effects of unilateral brain damage have been conducted. (See Goldstein, 1974, for a review.) Although there are a few exceptions, almost all have

Requests for reprints should be sent to Colin Martindale, Department of Psychology, University of Maine, Orono, Maine 04473.

shown that left-hemisphere damage leads to deficits on the verbal subtests of the Wechsler-Bellevue and the Wechsler Adult Intelligence Scale, whereas right-hemisphere damage leads to deficits on the Performance subscales (including Picture Arrangement) of these tests.

Dershowitz and Frankel (1975) used the slope of the regression line relating scaled score to rank of subtest difficulty (across all of their groups) as an index of degree of performance subtest deficit. The order of subtest performance from worst to best was as follows: Object Assembly, Picture Completion, Block Design, Picture Arrangement, Similarities, Arithmetic, Information, Comprehension. Thus, the steeper the slope, the greater the performance subtest deficit. Pooling the means presented in Dershowitz and Frankel's Table 1 (p. 127), I obtained slopes of .55 for 546 Jewish subjects, .31 for 30 Protestant subjects, and .25 for 94 presumably Catholic (Irish and Italian) subjects. (All of the slopes are positive because the tests were ordered on the basis of the performance of all groups in the first place.)

Several studies of unilateral brain damage include full tables giving mean subtest scaled scores of left- and right-hemisphere-damaged subjects on the Wechsler Adult Intelligence Scale and on the Wechsler-Bellevue Scale. If the Dershowitz and Frankel (1975) technique is applied to these means, slopes of .54 for 23 right-hemisphere-damaged subjects and of .06 for 21 left-hemisphere-damaged subjects are obtained for the means reported by Simpson and Vega (1971). Slopes of .35 for 31 right-hemisphere-damaged subjects and $-.33$ for 29 left-hemisphere-damaged subjects are obtained using the figures reported by Dennerll (1964). Unfortunately, in neither of these studies is the religious background of the subjects specified. The slope of .55 for Jewish subjects is essentially the same as one of the slopes obtained for right-hemisphere-damaged subjects and is somewhat higher than the other slope for right-hemisphere-damaged subjects. This is, of course, consistent with the hypothesis of a relative left-hemisphere dominance in Jewish subjects.

The notion that Jewish subjects exhibit a pattern of relative left-hemisphere dominance is not necessarily inconsistent with the hy-

pothesis that they tend to be field dependent. In fact, on the basis of a review of the effects of right-hemisphere damage, Pizzamiglio and Carli (1974) have hypothesized that right-hemisphere damage leads to field-dependent behavior. The implication of the hemispheric dominance hypothesis would be that the field-dependent test behavior of Jewish subjects does not arise from a lack of psychological differentiation on the level of personality but is, rather, an artifact of a more basic pattern of cognitive abilities. The hemispheric dominance hypothesis would seem, further, to be the more general and preferable one in that it can subsume the findings explained by the psychological differentiation hypothesis as well as others not explained by the latter hypothesis. Finally, the hemispheric dominance hypothesis predicts relative strengths as well as weaknesses for each of the religious groups.

In regard to intellectual test performance, the perceptual differentiation hypothesis cannot account for the poor performance of Jewish subjects on the WISC Picture Arrangement subtest, but the hemispheric dominance hypothesis can: Right-hemispheric damage depresses Picture Arrangement scores more than does left-hemispheric damage (McFie, 1960). It is not well established that Jewish subjects exhibit poor performance on the Embedded Figures Test. Only one of three studies obtained significant results in this direction. This may be because this test is dependent on both linguistic (left-hemisphere) and spatial (right-hemisphere) abilities: Aphasia causes decrements on this task (Teuber & Weinstein, 1956), but there is evidence that in patients without aphasia, right-hemisphere lesions cause greater decrements than do left-hemisphere ones (Russo & Vignolo, 1967). One study suggests that Jewish subjects perform more poorly than non-Jewish subjects on the Body Adjustment Test (Dershowitz, 1971). This is, of course, consistent with the psychological differentiation hypothesis. I know of no studies of the effects of lateralized lesions on this task, but the mediation of a wide variety of spatial orientation behaviors by the right hemisphere (cf. Luria, 1973) suggests that this finding should cause no problems for the hemispheric dominance hypothesis. Finally, the finding that

the figure drawings of Jewish children suggest field dependence (Dershowitz, 1971) is paralleled by the finding that right-hemisphere lesions produce a similar effect (Pizzamiglio, Note 2).

The hemispheric dominance hypothesis leads to a number of predictions that cannot be derived from the psychological differentiation hypothesis. On the basis of findings with lateralized lesions, we would expect Jewish subjects to perform worse than non-Jewish subjects on tasks such as three-dimensional size discrimination (cf. Weinstein, 1964), identification of faces (cf. Hécaen & Angelergues, 1963), and tactual form identification (cf. Pizzamiglio & Carli, 1974) but better than non-Jewish subjects on tasks involving verbal fluency and abstraction (cf. Luria, 1973).

The hypothesis that Jews, Protestants, and Catholics differ in degree of emphasis on processes dependent on one or the other hemisphere is consistent with a number of cultural and religious practices. For example, Jewish and Protestant religious rituals emphasize linguistic stimuli and deemphasize complex spatial stimuli such as images, stained glass windows, and elaborate architectural detail. Just the opposite is true of Catholic ritual. In old Hebrew, only consonants were written. Interestingly, the superiority of the left hemisphere in recognition of linguistic material is strongest for consonants (Studdert-Kennedy & Shankweiler, 1970). The right-to-left writing of Hebrew is also consistent with the finding that right-hemisphere weakness leads to perceptual left-side "neglect" (Luria, 1973). These and other consistencies do not, of course, prove the existence of the hypothesized hemispheric asymmetry nor do they tell us whether it causes the cultural differences or vice versa. (It seems likely that cultural differences bring about the differential emphasis on one or the other hemisphere, since Dershowitz and Frankel, 1975, presented evidence that the degree of Jewish weakness on spatial tasks decreases with acculturation as opposed to "traditional" subjects.) If the three groups do in fact tend to process and code information in the differential ways suggested, this would shed some potentially helpful light on the reasons for their historical conflicts. On a more prosaic level, it seems

clear that religion and ethnic background may be rather potent nuisance variables in use of intellectual test patterns in connection with diagnosis of left- and right-hemisphere brain damage.

Reference Notes

1. Litman, G. Personal communication to Z. Dershowitz, 1967.
2. Pizzamiglio, L. *Sophistication of the body concept in patients with unilateral brain damage*. Technical Report, Catholic University, Rome, Italy, 1973.

References

- Dennerll, R. D. Prediction of unilateral brain dysfunction using Wechsler test scores. *Journal of Consulting Psychology*, 1964, 28, 278-284.
- Dershowitz, Z. Jewish subcultural patterns and psychological differentiation. *International Journal of Psychology*, 1971, 6, 223-231.
- Dershowitz, Z., & Frankel, Y. Jewish culture and the WISC and WAIS test patterns. *Journal of Consulting and Clinical Psychology*, 1975, 43, 126-134.
- Goldstein, G. The use of clinical neuropsychological methods in the lateralization of brain lesions. In S. J. Dimond & J. G. Beaumont (Eds.), *Hemisphere function in the human brain*. New York: Halsted, 1974.
- Hécaen, H., & Angelergues, R. *La cécité psychique*. Paris: Masson, 1963.
- Luria, A. R. *The working brain*. New York: Basic Books, 1973.
- McFie, J. Psychological testing in clinical neurology. *Journal of Nervous and Mental Disease*, 1960, 131, 383-393.
- Pizzamiglio, L., & Carli, R. Visual, tactile and acoustic embedded figures test in patients with unilateral brain damage. *Cortex*, 1974, 10, 238-246.
- Russo, M., & Vignolo, L. A. Visual figure-ground discrimination in patients with unilateral cerebral disease. *Cortex*, 1967, 3, 114-127.
- Simpson, C. D., & Vega, A. Unilateral brain damage and patterns of age-corrected WAIS subtest scores. *Journal of Clinical Psychology*, 1971, 27, 204-208.
- Studdert-Kennedy, M., & Shankweiler, D. Hemispheric specialization for speech perception. *Journal of the Acoustic Society of America*, 1970, 48, 579-594.
- Teuber, H.-L., & Weinstein, S. Ability to discover hidden figures after cerebral lesions. *Archives of Neurology and Psychiatry*, 1956, 76, 369-379.
- Weinstein, S. Deficits concomitant with aphasia or lesions of either cerebral hemisphere. *Cortex*, 1964, 1, 154-169.
- Wendt, R. A., & Burwell, E. Test performance of Jewish day-school students. *Journal of Genetic Psychology*, 1964, 105, 99-103.
- Witkin, H. A. A cognitive style approach to cross-cultural research. *International Journal of Psychology*, 1967, 2, 233-250.

Received October 19, 1977 ■

Therapeutic Relationship in Behavior Therapy: An Empirical Analysis

Julian D. Ford
University of Delaware

The determinants and predictive utility of the client's perception of the therapeutic relationship (CPTR) were investigated in the context of a behavior therapy clinical research project evaluating three approaches to assertion training. Individual differences in therapists were a significant determinant of CPTR. Neither therapy type nor therapy session accounted for much variance in CPTR, although CPTR ratings by individual clients varied from session to session. CPTR was an effective predictor of dropping out when measured early in therapy, and of immediate posttherapy client gains when measured in a mid-to late therapy session, but not of long-term maintenance of client improvements. Patterns of therapist behavior that were predictive of CPTR at three time points in therapy are delineated. It is speculated that CPTR is largely a function of the degree to which the client's expectation of the therapist and the consequences of therapy are being fulfilled. It is concluded that CPTR has significant predictive value, and perhaps also causal impact, in behavior therapy.

The therapeutic relationship between client and therapist is widely acknowledged as a central factor in producing positive outcomes in psychotherapy and counseling (cf.

This study is based on a dissertation submitted in partial fulfillment of the requirements of the PhD degree at the State University of New York at Stony Brook. The guidance and encouragement of Leonard Krasner, Committee Chairman, and Marvin R. Goldfried, David Pomeranz, David Cross, and Aaron Carton, the other committee members, is gratefully acknowledged. Philip C. Kendall contributed very valuable comments on an earlier draft of the article. Special thanks are due to the persons who served as therapists in the study: Virginia Roswell, Judith Graeff, Lynn Turlette, Cheryl Fishbein, James Waters, Joan Broderick, Charles Harris, Gary Swan, Steve Friedman, Arthur Nezu, Arthur Stone, Jerry Friedman, and Leonidas Castro; and to Marvin Goldfried for enabling the study to be conducted in his clinical research project.

The study was conducted in conjunction with an assertion training outcome study reported by Linehan, Goldfried, and Goldfried (Note 1). Results concerning the comparative efficacy of the three therapy interventions are detailed in that report.

Requests for reprints should be sent to Julian D. Ford, Department of Psychology, Wolf Hall, University of Delaware, Newark, Delaware 19711.

Howard & Orlinsky, 1972; Kiesler, 1973; Marsden, 1971; Meltzoff & Kornreich, 1970; Mitchell, Bozarth, & Krauft, 1977). Within the behavior therapy framework, increasingly greater attention has been paid to the therapeutic relationship in recent years (cf. Wilson & Evans, 1976). However, very little empirical research has been reported to directly evaluate the impact of variations in the therapeutic relationship on the outcome of behavior therapies. This study represents an initial attempt to address that issue.

Although the therapeutic relationship can be assessed from several vantage points (i.e., the client, the therapist, experienced clinician observers, nonexpert observers), and in terms of either specific client-therapist behavioral interactions or the global experiencing of the client and therapist, only one approach to the measurement of the therapeutic relationship has been well validated as a predictor of therapy outcome: the client's perception of the therapeutic relationship (CPTR) (cf. Gurman, 1977). Despite optimistic early reviews (Truax & Mitchell, 1971), there is no consistent evidence that ratings by nonexpert observers of the therapeutic relationship are predictive of outcome in therapy (cf. Mit-

chell et al., 1977). All other approaches to the measurement of therapy processes have received little, if any, empirical evaluation as predictors of therapy outcome, with very few promising results (e.g., Dietzel & Abeles, 1975; Mintz & Luborsky, 1971; Rice, 1965).

Although CPTR has been found to be an excellent predictor of therapeutic outcome in client-centered, psychoanalytic, vocational guidance, and "personal" therapy and counseling (Gurman, 1977), this has been done only three times in a behavior therapy context, and these studies have had important methodological limitations. In an analogue desensitization study conducted with mildly phobic volunteer subjects, Carmichael (cited in Gurman, 1977) found CPTR to be inversely predictive of decreases in avoidance behavior and not predictive of changes in subjective fear. Ryan and Gizynski (1971) found that retrospective CPTR reports were correlated with therapists' global ratings of client improvement after receiving unspecified types of therapy by self-identified behavior therapists. Sloane, Staples, Cristol, Yorkston, and Whipple (1975) also showed that retrospective CPTR reports were predictive of outcome in unspecified types of therapy administered by self-identified behavior therapists, using global ratings by clients, therapists, and expert judges as their outcome criteria.

The designs and results of these studies leave a major question unanswered. Is CPTR predictive of changes in therapy clients' *specific behaviors and self-perceptions* that are caused by participation in a *clearly described naturalistic* program of behavior therapy? Although the value of CPTR as a predictor of global improvement in behavior therapy has been demonstrated, it is unclear whether it is predictive of the more behaviorally specific criteria that are essential in behavior therapy. And, the therapy types used in these studies were either too incompletely described to permit replication or too artificial to be generalized to actual clinical practice.

Furthermore, these studies all measured CPTR retrospectively, so it is not possible to infer whether it would, if assessed prior

to the termination of therapy, serve as a clinically viable predictor of therapy outcome. Nor can we say whether CPTR remains constant across the course of therapy or whether it is differentially effective as a predictor at different time points.

Finally, there has been very little research in any therapy approach, and none in a behavior therapy context, that delineates the determinants of CPTR (cf. Gurman, 1977). If CPTR is important, how does it come about? Research on client and therapist personal and professional characteristics (e.g., age, psychological status, expertise, mood, dogmatism) as predictors of CPTR is sparse and inconclusive at present (see Gurman, 1977). When observers' ratings of therapist in-session behavior are tested as predictors, many verbal, vocal, kinesic, and proxemic behaviors have been examined, but very few have emerged even in one study (let alone in replications) as viable predictors of CPTR (e.g., therapist interruptions; vocal expression of concern and involvement; eye contact, forward trunk lean; see Gurman, 1977). Clearly, very little is known about the determinants of CPTR in any therapy context.

One study that has examined this issue using a factorial analysis of variance design demonstrated that the emergent events in each therapy session were the major causal factor underlying CPTR (Howard, Orlinsky, & Perlstein, 1976). But what are those events (e.g., What is it that the therapist is saying or doing that facilitates positive CPTR?), and does this hold true in a behavior therapy context as well? These issues are examined in this study.

Method

Clients

The clients were 39 volunteers from scattered suburban communities in the vicinity of a major east coast university, who responded to announcements in the media of a time-limited therapy program for assertion training. Persons responding to the announcements were sent a brief description of the program, which included the following information:

We hope to achieve two goals: first of all, to assist people like yourself who have difficulty

acting assertively or who experience anxiety or nervousness in situations where they want to act assertively. At the same time, we hope to obtain valuable information regarding the types of treatment methods that are most effective for different individuals. All of the treatment procedures we will be using in the assertion program have been shown to be effective for some individuals.

The therapists will be PhD candidates in clinical psychology trained and supervised by Dr. Goldfried. . . . We are unable to accept anyone into the program who is currently being seen in psychotherapy. . . . The treatment sessions will consist of weekly *individual* sessions carried out over a 2-month period, with each session lasting approximately 1 hour. . . . Because the program is being funded by the National Institute of Mental Health, there will be no charge to you. . . . We want to again emphasize that your questionnaire responses and what transpires during the actual assertive training sessions will be kept strictly confidential.

Persons applying to the program were screened (a) to eliminate client sex as a factor, and since the vast majority of applicants were women, only females were accepted as clients; (b) university students were not accepted in order to obtain a representative sample of community adults; (c) persons concurrently undergoing psychotherapy were not accepted, nor were persons who showed evidence of severe marital problems or thought disorder; and (d) only applicants who scored below zero on the Rathus (1973) Assertiveness Schedule and above three on an assertion screening inventory designed especially for this study (Personal Reaction Inventory, Form D) were accepted as participants. All persons who volunteered were screened in an interview conducted by an experienced clinical psychologist. Applicants who met the screening criteria were accepted as clients until all therapists participating in the study were matched with three clients; thereafter, applicants were referred to other clinics and assertion training programs in the area with the explanation that the program could no longer accept further applications.

All clients reported in the screening interview that problems with assertiveness had generalized debilitating effects on their lives (e.g., self-confidence, marriages, vocations). The average pretherapy score on the Rathus (1973) Assertiveness Schedule was -46.3 , markedly below the mean scores reported by Rathus for a nonclinical college population (i.e., $.3$ and 1.6). Thus although it must be emphasized that these clients were solicited for assertion training therapy rather than self-referred for psychotherapy, there is evidence that the problems in living that confronted them were clinically serious.

The clients' average age was 36.5 years (range = 22-60), and they had an average education level of

2 years in college (range = 10-17 years of education). Seventy percent were married. Ten were unemployed outside the home, 8 had part-time jobs, and 12 held full-time jobs.

Therapists

The 13 therapists were advanced graduate students in a clinical psychology training program (M age = 28.6). Eight were male, and 5 were female. All but 1 were inexperienced (Mdn = seven prior cases).

Therapy Types

Clients were matched using within-group procedures on the basis of age and scores on all pretest measures (see Table 1) and were then randomly assigned to one of three therapy types (Linehan, Goldfried, & Goldfried, Note 1): interactive, in which therapists focused on providing nondirective support to the client; instigation, in which homework assignment to engage in assertive behaviors was provided as a supplement to therapist support; and, rational restructuring, in which clients were trained to restructure their cognitive self-statements (cf. Ellis, 1970) through modeling and behavior rehearsal, in addition to receiving behavioral instigation and emotional support. All clients received an explanation of why people are unassertive and how one can become more assertive that was congruent with their therapy types in the initial session.

Each therapist conducted each type of therapy with an individual client over eight 1-hour weekly sessions. Two therapists had one client drop out, one therapist had two clients drop out, and one therapist had one client drop out and two clients who failed to provide pretest data. Thus, there were 32 clients who completed the entire therapy process.¹

Measurement of the Perceived Therapeutic Relationships

The Relationship Inventory Form G (RI-G), as developed by Gorman (1973), was used to assess the client's perception of the therapeutic relationship at three points in the therapy process (i.e., immediately after the third, sixth, and final therapy sessions). The RI-G is a 30-item true-false written questionnaire that was derived from similar previous instruments. Ten-item subscales can be scored for empathy, warmth, and genuineness, but reli-

¹One of the clients who completed therapy was not included in the data analyses because her therapist had both of his other clients drop out, and was thus not representative of the other therapists (none of whom had more than one client drop out).

ability has been assessed only for the entire scale; split-half Spearman-Brown reliability was reported to be .86 by Gurman (1973).

Measurement of Therapist Behaviors

Twelve 3-minute samples were taken from audiotapes of each of the 31 complete therapy cases.

Table 1
Outcome Measures

Variables measured pre, post and follow-up	
Rathus Assertiveness Scale (Rathus, 1973)	
Personal Reaction Inventory Form D (Linehan, Goldfried, & Goldfried, Note 1)	
Self-Esteem Inventory (Robinson & Shaver, 1973)	
Fear of Negative Evaluative Scale	
Social Avoidance/Distress Scale (Watson & Friend, 1969)	
Bem Sex-Role Inventory (Bem, 1974)	
Masculine	
Feminine	
Androgyny	
Multiple Affect Adjective	
Checklist (Zuckerman & Lubin, 1965)	
Anxiety	
Depression	
Hostility	
S-R Inventory of Anxiety (Endler & Hunt, 1968)	
Refusal situations	
Initiation situations	
Total score	
S-R Inventory of Hostility (Endler & Hunt, 1968)	
Refusal situations	
Initiation situations	
Total score	

Variables measured pre and post only

Role-playing self-ratings
Nervousness
Baseline nervousness
Anger
Guilt
Role-playing observer-rated behavior
Eye contact duration
Loudness/affect (first response)
Loudness/affect (M of all responses)
Speech disfluency (first response)
Speech disfluency (M)
Content: assertiveness level
No. challenges by confederate

Variables measured post only

Stress test observer-rated behavior
Content: assertiveness level
No. challenges by confederate
(Friedman, 1971)

Therapy Sessions 1, 3, 6, and 8 were sampled, with one 3-minute sample taken from the middle of the first (i.e., early), second (i.e., middle), and final (i.e., late) third of each session. Eight trained research assistants rated the therapist's behavior from the audiotaped time samples and written transcripts. Each rater rated an equal number of tape segments from all therapists, clients, therapy types, therapy sessions, and in-session time periods. Raters were assigned to teams of two or three, and each team rated one third of the behavior categories so as to minimize rater overload.

The therapist behavior code that was used for ratings consists of 48 behavior categories, each operationally defined in detail and derived from existing therapy process measurement instruments (see Table 2). Each category was rated once every 20 sec, or nine times in each 3-min sample. The data were condensed by means of arithmetic averaging to provide one score for each 3-min time sample.

Reliability was assessed by having each rater re-rate 48 (for the three-person teams) or 72 (for the two-person team) 3-minute segments already rated by each other rater in his/her team. These segments were equally distributed across therapists, clients, therapy types, therapy sessions, and in-session time periods. Reliability was assessed by means of Pearson product-moment correlations for variables that were scored as frequency counts or ordinal ratings. Reliability coefficients ranged from .70 to .99, with a median of .82. For the variables that were scored simply as occurrence or nonoccurrence, Cohen's (1960) Kappa was used as a reliability estimate, since it accounts for chance agreements when the more typically used percentage agreement score does not. When Kappa was calculated for interrater agreement on both occurrence and nonoccurrence, reliability ranged from .77 to 1.00, with an average of .87. With the much more conservative test of agreements on occurrences only, the levels were .40 or higher (i.e., well above chance), with only one exception (i.e., topic change, .28), and with an average of .73.

Measurement of Therapeutic Outcome

Thirty separate variables were used to assess therapy outcome (see Table 1). Seventeen variables were derived from written questionnaires completed by clients before, immediately after, and 2 months after therapy. These self-report measures assess a variety of components of assertion, including self-perception of assertiveness, anxiety, hostility, depression, and masculinity-femininity as general traits and as behavioral reactions to specified situations. All of the measures have been well researched and have been extensively used to evaluate outcome in prior assertion training research (cf. Rich & Schroeder, 1976), although none have clearly demonstrated external validity.

Eleven variables were taken from the client's

Table 2
Therapist Behavior Code Categories

Frequency count categories

No. therapist utterances
Verbal productivity
Encouragements to continue or complete utterance
Interruptions

Ordinal rating categories

Nonverbal
warmth
certainty and sincerity
control and relaxation
energy
emotional responsivity

Occurrence-nonoccurrence categories

Laugh (during client utterance)
Informational statement
fact
possibility
Request information
closed ended
open ended
Request Action
command
suggestion
Contentless communication
Simultaneous speech
Implements technique
Models assertive thoughts
Reflects client's feelings/fears
Implied question or statement
Restatement
Agreement
Disagreement
Client strength emphasized
Client weakness emphasized
Topic change
Interpretation based on client's viewpoint
Interpretation based on therapist's viewpoint
Reassurance
Clarification
Disfluency
Filled pause
Silent pause
First person singular ("I" statement)
First person plural ("We" statement)
Content
Emotions
Cognitions
Behavior
Client
Therapist
Client's significant others
Past
Present
Future
In session
Outside session

subjective emotional reactions to, and observer ratings of client behavior in, six role-playing simulations requiring assertion and were measured only at pretest and posttest (Table 1).

Finally, two variables were assessed at posttest only: observer ratings of 2 categories of client behavior in an analogue "stress test."

Trained research assistants served as observers for the behavioral outcome measures. Interrater agreement levels were calculated by product-moment correlations between two observers for each variable. Reliability levels were as follows: role-playing assertiveness content ($M = .85$, range = .84-.87); role-playing loudness/affect ($M = .92$, range = .78-.97); role-playing disfluency ($M = .88$, range = .83-.92); and stress test assertiveness content ($M = .84$, range = .69-.96).

Pretest, posttest, and follow-up scores were converted to outcome scores through the use of residual gain scores (Manning & DuBois, 1962), which were used rather than raw difference scores because of their greater reliability and their control for regression effects due to pre-post or pre-follow-up correlation.²

Results

Variation in the Perceived Therapeutic Relationship

Three two-way analyses of variance (ANOVAS) were conducted, with RI-G scores as the dependent measure, to determine the proportions of variance in CPTR accounted for by differences among therapists, therapy types, and therapy sessions: (a) Therapists \times Therapy Types (pooling over therapy sessions); (b) Therapists \times Therapy Sessions (pooling over therapy types); and (c) Therapy Types \times Therapy Sessions (pooling over therapists). A single three-way ANOVA was not used, since it would have involved one data score per cell, thus confounding the Therapist \times Therapy Type and Therapist \times Therapy Type \times Therapy Sessions interaction terms with the error term. Variance proportions were calculated based on expected mean square formulas (Winer, 1971) and the assumption that therapist is a random between-subjects factor, therapy type is a fixed between-subjects factor, and therapy session is a fixed within-subjects (repeated measures) factor (cf. Endler & Hunt, 1968).

² Those correlations were noteworthy in this study, with 75% falling in the .32-.78 range ($p < .05$) and more than half greater than .52.

Table 3
Determinants of Variance in the Perceived Therapeutic Relationship

Factor	A × B	A × C	B × C	Overall
Therapist (A)	.20	.43	—	.315
Therapist type (B)	.02	—	.01	.015
Therapy session (C)	—	.08	.07	.075
A × B	.18	—	—	.180
A × C	—	.05	—	.050
B × C	—	—	.07	.070
Residual	.60	.44	.85	.295

Note. Determinants are expressed as proportions, varying between .00 and 1.00. Overall variance proportions for therapist, therapy type, and therapy session were determined by numerically averaging the two values for each. The overall variance proportion for the residual term was calculated by subtracting the overall variance proportions of all other factors from 1.00. The residual term includes the variance due to the highest order interaction term and error.

The results (Table 3) demonstrate that differences between the individual therapists had a significant effect on CPTR, the interaction of therapists and therapy types had a moderate impact, and all other factors exerted only a minimal influence. Only one factor achieved a significant *F* in any ANOVA (i.e., Therapist, in the Therapist × Therapy Session ANOVA).

A *t* test was conducted to examine whether the therapist's gender was a significant determinant of CPTR ratings. The results showed that male and female therapists did not receive significantly different CPTR ratings, $t(91) = .1$, and this is reflected in the mean ratings for males (27.1) and females (26.8).

Perceived Therapeutic Relationship as a Predictor of Outcome

Correlation coefficients were calculated between RI-G scores and the scores for each outcome measure separately for Sessions 3, 6, and 8. The results (Table 4) show that RI-G scores were not consistently predictive of therapy outcome at Session 3, with only 4 of 47 correlations at or near significant. However, Session 6 RI-G scores were consistently predictive of posttest outcome, with 18 of 30 correlations at or near (i.e., $p = .06$) significant, all in the expected direction. Session 6 RI-G scores were not consistently predictive of follow-up outcome, with only 4 of 17 correlations significant. Session 8

RI-G scores were not consistently predictive of outcome, with only 4 of 47 correlations significant.

CPTR was assessed as a predictor of premature termination by comparing the Session 3 RI-G scores of the five dropouts with those of the 31 full-term clients. Mean RI-G levels for the two groups were 14.6 and 27.3, respectively. A *t* test showed that the clients who completed therapy gave significantly higher RI-G ratings for Session 3 than did the dropouts, $t(34) = 8.1$, $p < .001$. Few false positives or negatives were found, as well: RI-G scores for three dropouts were much lower than those for any full-term client; the fourth dropout's RI-G score was equal to the lowest RI-G score for any full-term client; and, the fifth dropout's RI-G score was lower than all but two of the full-term clients' RI-Gs.

Therapist Behavior as a Predictor of CPTR

The therapist behavior categories were grouped into 14 composite behavioral "styles" on an a priori basis to permit for relatively reliable stepwise multiple regression analyses with the small sample size (see Table 5). The multiple regression analyses were conducted separately for each of the three therapy sessions for which RI-G ratings were obtained.

Results from the nine multiple regression analyses are presented in Table 6. Overall, therapist behaviors did not account for the

major part of the variance in RI-G scores, generally explaining between 15% and 30% of the variance in CPTR. However, in the early phase of both Sessions 6 and 8, therapist behavior accounted for substantial proportions of RI-G variance (i.e., 48% and 53%).

Although there was much variation from session to session, and time period to time period, certain styles were consistently positively related to favorable CPTR (i.e., focus on behavior or cognitions; nonverbal style; cognitive restructuring), whereas one was often inversely related to RI-G scores (i.e., significant others).

Discussion

The fact that the client's perception of the therapeutic relationship seems to be a potent factor in behavior therapy is consistent not only with prior research (Gurman, 1977) but also with two important current developments in behavior therapy. First, there is growing recognition that the (ostensibly) same technique (e.g., desensitization), when applied by different therapists or at different time points or in different settings or with different clients, may be implemented quite differently and with different outcomes. Behavior therapists have

Table 4
Therapeutic Relationship as a Predictor of Outcome

Outcome measure	RI-G session		
	3	6	8
Posttest outcome			
Rathus Assertiveness Schedule		.29*	
Fear of negative evaluation		-.32**	
Social avoidance/distress		-.33**	
MAACL anxiety		-.39**	
Bem SRI: Masculine		.29*	
S-RIA Initiation		-.45***	
S-RIA Refusal		-.53***	
S-RIH Initiation		-.43***	
S-RIH Refusal		-.50***	
S-RIA Total		-.51***	
S-RIH Total		-.48***	
Role play			
Nervousness			
Baseline nervous	-.34**	-.52***	
Anger	-.32**	-.34***	
Eye contact		-.52***	
Disfluency (1st)		.32**	
Content		-.38**	
Challenges	-.39**	.28*	
	-.32**	.29**	
Follow-up outcome			
Fear of negative evaluation			
MAACL Anxiety		-.31**	
MAACL Depression		-.43***	
S-RIH Total		-.52***	
Bem SRI: Feminine		-.33**	
Bem SRI: Masculine			-.56***
S-RIA Initiation			-.46***
S-RIA Refusal			-.57***
			-.55***

Note. RI-G = Relationship Inventory Form G; MAACL = Multiple Affect Adjective Check List; SRI = Sex-Role Inventory; S-RIA = S-R Inventory of Anxiety; S-RIH = S-R Inventory of Hostility.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 5
Composite Therapist Behavior Variables

New variable	Combination of original therapist behavior variables
Nonverbal Voice quality	Nonverbal warmth + nonverbal certainty + nonverbal control + nonverbal energy + nonverbal responsivity - simultaneous speech - disfluency - filled pause - silent pause
Supportiveness	Encourages to continue or complete + agreement + emphasis on client strengths + reassurance - interruptions - disagreement - emphasis on client weakness
Instigation/prescription	Information statement: fact + information statement: possibility + request action: command + request action: suggestion + interpretation: client view + interpretation: therapist view + content: client + content: future
Assessment	Request information: closed ended + request information open ended + content: client + content: past + content: outside session
Nondirectiveness	Contentless communication + restatement - number of utterances - verbal productivity - implied question or statement
Technique implementation	Implements technique + models assertive thoughts + reflects client fears
Self-disclosure	Content: therapist
Here-and-now	Content: present + content: in session
Affect	Content: emotions
Cognitions	Content: cognitions
Behavior	Content: behavior
"I" statement	First-person singular
"We" statement	First-person plural
Significant others	Content: significant others

developed precautions (e.g., Paul, 1969) against the "therapist uniformity myth" (Kiesler, 1966) for many years now. However, it is recognized that it is not sufficient to only control for "nonspecific" effects: We must attempt to determine what it is in the moment-to-moment behaviors of therapists that affects clients and how these specific behaviors interact with behavior change techniques—and differences in therapists, clients, settings, and target objectives—to produce more or less successful outcomes (cf. Wilson & Evans, 1976). Only then will systematic training and research replication be possible.

Second, the significance of the client's perceptions and cognitions (i.e., attributions, self-statements, expectations) in behavior change and maintenance has been convincingly established by the cognitive therapists (e.g., Beck, 1970; Ellis, 1970) and cognitive behavior therapists (e.g., Lick & Bootzin, 1975; Mahoney, 1974). Self-perception,

self-reinforcement functions, expectancy of gain, cognitive distortions, and cognitive coping strategies have been extensively considered in recent behavior therapy theory and research. Now, CPTR can be added as a researchable cognitive factor of potential practical importance.

Variance in CPTR

Individual differences among therapists emerged as the primary determinant of CPTR, consistent with previous research on therapeutic process in contexts other than behavior therapy (cf. Meltzoff & Kornreich, 1970). Further research will be needed to pinpoint just what it is that differentiates therapists who evoke more or less CPTR (e.g., mood states; cognitive styles). A *t* test conducted to explore the possibility of sex differences was nonsignificant, indicating that therapist sex was not a factor in this study. It should also be noted that replication with

Table 6: *Therapist Style as a Predictor of the Client's Perception of the Therapeutic Relationship*

Predictor	<i>R</i>	<i>R</i> ²	<i>R</i> ² increase	Coefficient
Session 3 Time 2				
Behavior	.46	.21	.21	9.8
Significant others	.61	.37	.16	-7.6
Nonverbal quality	.38	.15	.15	2.1
Significant others	.51	.25	.10	-6.4
Session 3, Time 3				
"We" statements	.39	.15	.15	9.1
Instigation	.49	.25	.10	12.5
"We" statements	.39	.15	.15	9.1
Nonverbal quality	.47	.22	.07	1.2
Session 6, Time 1				
Cognitions	.41	.17	.17	13.4
Significant others	.59	.35	.18	-10.1
Here and now	.69	.48	.13	-8.1
Session 6, Time 2				
Emotions	.45	.20	.20	9.05
Session 8, Time 1				
Behavior	.40	.16	.16	19.1
Significant others	.59	.35	.19	-11.7
Supportiveness	.67	.45	.10	-24.6
Nondirectiveness	.73	.53	.08	24.9
Session 8, Time 2				
Cognition	.36	.13	.13	8.0
Session 8, Time 3				
"I" statements	.36	.13	.13	7.3

Note. There were no predictors for Session 3, Time 1 and Session 6, Time 3.

more experienced therapists will be necessary to rule out the possibility that individual differences among therapists are only factors with relatively inexperienced therapists.

Surprisingly, therapy type had virtually no effect on CPTR. Apparently, the same therapist can establish rapport as effectively when applying systematic cognitive and behavioral interventions as when focusing exclusively on offering nondirective emotional support. This result is consistent with prior research showing that behavior therapists are equally able to evoke positive CPTR as compared to other therapists of other orientations (e.g., Fischer, Paveza, Kickert, Hubbard, & Grayston, 1975; Sloane et al., 1975).

The particular therapy session also accounted for relatively little of the variance in CPTR. However, in contrast to previous research, which showed that CPTR measures have excellent temporal stability (Gurman, 1977), the correlations between RI-G

scores at the three different time points were .05 (Session 3 to Session 6), .05 (Session 3 to Session 8), and -.56 (Session 6 to Session 8). The data suggest that CPTR should be measured at several time periods across the course of therapy when used as either a research or clinical tool.

Because only one *F* ratio achieved significance in the three ANOVAs, and due to the relatively small sample sizes, replication with larger *N*s seems advisable to confirm the finding that only therapist factors contributed substantially to variance in CPTR, and to rule out the possibility of chance findings.

CPTR as a Predictor of Outcome

CPTR was shown to have value as a predictor of two key therapeutic outcomes: staying in therapy and making changes in behavior and self-perception. Although this is consistent with previous research (e.g., Ryan

& Gizynski, 1971; Sloane et al., 1975), replication is certainly needed with other behavior therapies, more experienced therapists, clinic-referred rather than volunteer clients, longer term and open-ended therapies, and in a naturalistic rather than research context before this conclusion can be generalized beyond the present study's context. However, the results have good external validity for several significant settings, populations, and therapy types, including novice therapists-in-training, brief time-limited therapies (Watzlawick, Weakland, & Fisch, 1974), therapeutic and educational interventions in which assertion training is the primary vehicle and target (Rich & Schroeder, 1976), and most behavior therapy outcome research projects (O'Leary & Wilson, 1975).

CPTR was never an effective predictor of long-term maintenance of client improvements as measured at the 2-month follow-up. This suggests that CPTR facilitates and/or is predictive of short-term therapeutic change but that long-term behavioral maintenance requires special therapeutic intervention (e.g., training in self-management skills; restructuring the client's natural environment). This is consistent with a major principle of behavior therapy: Behavioral generalization and maintenance must be programmed rather than lamented (cf. Baer, Wolf, & Risley, 1968). It also reconfirms the notion that therapist facilitativeness is a necessary but *not* sufficient ingredient in effective counseling and psychotherapy.

Therapist Behavior as a Predictor of CPTR

Given the importance of positive CPTR, the critical question remains as to what a therapist can do to establish rapport with each client. (i.e., What is the RI-G really measuring?) In Session 3, there were no effective predictors from the first in-session time period, but in the middle part of the session a focus on behavior or a warm, relaxed, and energetic nonverbal style, plus a focus on the client and not on her significant others, was optimal for CPTR. In the final time period, a stress on a collegial client-therapist relationship (i.e., "We" statements) plus either encouragement to take action or

a positive nonverbal style were associated with higher RI-G scores. Thus, it seems that CPTR is maximized in this early therapy session when the therapist communicates verbal and/or nonverbal encouragement, involvement, concern, and respect for the client. The behavioral correlates of CPTR in this session closely parallel those identified by previous research (Gurman, 1977). This is not surprising, since much of the previous research was done using behavior samples from sessions early in therapy. This type of verbal and nonverbal approach is thought to be essential to creating "rapport" in early sessions by behavior therapists (e.g., Wilson & Evans, 1976) as well as by therapists of other theoretical orientations.

In the early phase of Session 6, a focus on the client rather than on her significant others, and in particular on her cognitive responses to extratherapy session events and situations, was related to higher CPTR ratings. As in Session 3, CPTR in Session 6 is maximized when the therapist shows particular interest in the client (and minimized when the focus is on her significant others), but more specifically on her extratherapy cognitive reactions and coping strategies. Starting off on the right foot seems important in Session 6, because therapist behavior is relatively insignificant as a predictor of CPTR in the middle and late phases of this session.

Similarly, in the final therapy session, therapist behavior is strongly related to RI-G ratings only in the early phase. At that point, a focus on the client and not on her significant others, with the therapist talking relatively little and largely reflecting back the client's messages, and with little direct support explicitly communicated to the client, were predictive of higher RI-G ratings. At the middle time point, a focus on cognitions was optimal, and in the final part of the final session, "I" statements were the sole predictor of positive CPTR. This suggests that therapists enhanced CPTR in Session 8 by taking actions that facilitated a smooth and, for the client, complete transition between therapy and termination: allowing the client to take charge and wrap-up any un-

finished business that she might have; not attempting to belatedly bolster the client's self-perceptions through supportive statements; dealing with the client's expectations, concerns, and fears; and providing personalized feedback to the client.

Thus, consistent with past research in a nonbehavioral therapy context (e.g., Barrington, 1961), the pattern of therapist behaviors that emerged as predictive of CPTR was not static; it changed systematically from session to session. Clearly, therapists must tailor their style of interaction to fit the changing goals and needs of the client at different time points in therapy.

These data suggest a potential explanation for why clients' RI-G ratings varied from session to session, and why RI-G scores were differentially effective as predictors of different outcomes at different time points in therapy. The data do not support the possibility that clients were basing their RI-G ratings on the same therapist behaviors in each session and that their therapists simply behaved very differently in different sessions: With one exception (i.e., a focus on significant others was detrimental to CPTR in all sessions), RI-G ratings were associated with different patterns of therapist behaviors in each session and at different time periods.

A more plausible explanation is that clients based their RI-G ratings on how well the therapist and the therapy were fulfilling their expectations, and that different factors influenced the clients' expectations at different points in therapy. Early in therapy, a global good-bad therapist/therapy dimension that has been hypothesized based on empirical data by previous researchers (cf. Mitchell et al., 1977), and that has little to do with subsequent client gains, but that serves as a basis for the client to either stay in or drop out of therapy seemed to be most important—hence, the importance of such behaviors as nonverbal style and emphasis on a collegial client-therapist relationship at this point. At the mid to late portion of therapy, CPTR seemed to be largely a function of the gains that the client saw herself making toward her objectives (i.e., becoming more assertive and less anxious in extra

therapy social situations). The therapists seemed to enhance, and perhaps elicit, these positive self-perceptions by encouraging the client to discuss her successful attempts at assertion from the past week and the cognitions that accompanied such occurrences. Finally, in the last therapy session, RI-G scores appeared to assess the client's feelings about ending therapy. Those who felt finished and ready to move ahead on their own gave high ratings, whereas those who felt unready to end the supportive relationship expressed this fear and dissatisfaction through low RI-G scores. Therapist actions that prevented the latter and facilitated the former cognitive set (e.g., allowing the client to deal with the issues that she feels are essential, giving meaningful feedback) enhanced CPTR. The significant inverse relationship between RI-G scores for Sessions 6 and 8 seems to have occurred because some of the clients who experienced significant gains in therapy (and thus tended to give high RI-G ratings during Session 6) were nevertheless not ready to terminate therapy and therefore gave lower RI-G ratings in Session 8. Conversely, some of the clients who did not make significant gains (and who gave low CPTR ratings in Session 6) were simply insufficiently assertive to resist the strong "hello-and-goodbye" demand characteristics that generally accompany the end of therapy or simply adjusted their expectations and self-perceptions so that despite failing to make significant changes, they did indeed feel ready to leave therapy. A scan of the data for individual clients supported this explanation: 80% of the clients increased or decreased their RI-G ratings from Session 6 to Session 8, and they were relatively evenly distributed in four groups paralleling those hypothesized (i.e., some clients who made significant gains increased their RI-G ratings while some lowered them; similarly, some clients who failed to improve significantly in therapy raised their RI-G ratings and others lowered them).

Clearly, more detailed research must evaluate these hypotheses before they can be considered as more than speculative. In particular, a more direct assessment of the factors that are hypothesized as bases for the client's

RI-G ratings at different points in therapy is necessary (e.g., in addition to administering the RI-G, clients could be asked specifically about their global evaluations of the therapist and therapy, their perceptions that they were benefiting from therapy, and their readiness to end therapy). Further refinement of the therapist behavior code is another important future direction to eliminate the need for a priori groupings of the separate therapist behavior code variables. The measures that contributed most to the prediction of CPTR would seem to be excellent foci for such research (e.g., significant others, non-verbal quality, we statements, cognitions). Also, other potential determinants of CPTR must be examined in light of the finding that therapist behavior rarely accounted for as much as 50% of the variance in RI-G scores (e.g., client expectations and self-perceptions; the nature of the client-therapist interaction; and therapist characteristics).

To conclude, this study has demonstrated that CPTR is an effective predictor of dropping out and *short-term* client gains in a behavior therapy context. Differences among individual therapists contributed significantly to variance in CPTR, but therapy type and session did not. Further, a set of specific therapist behaviors, which were remarkably similar to those postulated by Rogerian clinicians (Mitchell et al., 1977), were found to be predictive of positive CPTR. The limitations of this study mandate its replication before the results can be considered as definitive, yet the importance of the therapeutic relationship in behavior therapy seems clear. However, it must be recalled that CPTR was never an effective predictor of the fundamental criterion of therapeutic effectiveness: *long-term* client gains. We cannot rule out the possibility that in view of the potential reactivity of the self-report and role-playing measure of outcome (Ciminero, Calhoun, & Adams, 1977), the strong demand characteristics implied by a facilitative therapist may have been responsible for the pre-post client gains that were related to CPTR. The fact that several outcome measures showed a correlation with CPTR militates against but does not rule out this hypothesis. CPTR thus appears to facilitate the process of change in

behavior therapy, but it is clearly not a sufficient basis for fully effective intervention.

Reference Note

1. Linehan, M., Goldfried, M. R., & Goldfried, A. P. *A systematic evaluation of three approaches to assertion training*. Unpublished manuscript, State University of New York at Stony Brook, 1977.

References

- Baer, D. M., Wolf, M. M., & Risley, T. Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1968, 1, 91-97.
- Barrington, B. L. Prediction from counselor behavior of client perception and case outcome. *Journal of Counseling Psychology*, 1961, 8, 37-42.
- Beck, A. T. Cognitive therapy: Nature and relation to behavior therapy. *Behavior Therapy*, 1970, 1, 184-200.
- Bem, S. L. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1974, 42, 155-162.
- Ciminero, A., Calhoun, K., & Adams, H. (Eds.). *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, 20, 37-46.
- Dietzel, D. S., & Abeles, N. Client-therapist complementarity and therapeutic outcome. *Journal of Counseling Psychology*, 1975, 22, 264-272.
- Ellis, A. *The essence of rational therapy*. New York: Institute for Rational Living, 1970.
- Endler, N. S., & Hunt, J. McV. S-R inventories of hostility and comparisons of the proportions of variance from persons, responses, and situations for hostility and anxiousness. *Journal of Personality and Social Psychology*, 1968, 9, 309-315.
- Fischer, J., Paveza, G. J., Kickert, N. S., Hubbard, L. J., & Grayston, S. B. The relationship between theoretical orientation and therapists' empathy, warmth, and genuineness. *Journal of Counseling Psychology*, 1975, 22, 399-403.
- Friedman, P. H. The effects of modeling and role playing on assertive behavior. In R. D. Rubin, H. Fensterheim, A. A. Lazarus, & C. M. Franks (Eds.), *Advances in behavior therapy*. New York: Academic Press, 1971.
- Gurman, A. S. Effects of therapist and patient mood on the therapeutic functioning of high- and low-facilitative therapists. *Journal of Consulting and Clinical Psychology*, 1973, 40, 48-58.
- Gurman, A. S. The patient's perception of the therapeutic relationship. In A. S. Gurman & A. M. Razin (Eds.), *Effective psychotherapy: A handbook*. New York: Pergamon Press, 1977.
- Howard, K. I., & Orlinsky, D. E. Psychotherapeutic processes. *Annual Review of Psychology*, 1972, 23, 615-668.

- Howard, K. I., Orlinsky, D. E., & Perlstein, J. Contribution of therapists to patients' experiences in psychotherapy: A components of variance model for analyzing process data. *Journal of Consulting and Clinical Psychology*, 1976, 44, 520-526.
- Kiesler, D. J. Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin*, 1966, 65, 110-136.
- Kiesler, D. J. *The process of psychotherapy: Empirical foundations and systems of analysis*. Chicago: Aldine, 1973.
- Lick, J., & Bootzin, R. Expectancy factors in the treatment of fear: Methodological and theoretical issues. *Psychological Bulletin*, 1975, 82, 917-931.
- Mahoney, M. J. *Cognition and behavior modification*. Cambridge, Mass.: Ballinger, 1974.
- Manning, W. H., & DuBois, P. H. Correlational methods in research on human learning. *Perceptual and Motor Skills*, 1962, 15, 287-321.
- Marsden, G. Content analysis studies of psychotherapy: 1954 through 1968. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis*. New York: Wiley, 1971.
- Meltzoff, J., & Kornreich, M. *Research in psychotherapy*. New York: Atherton Press, 1970.
- Mintz, J., & Luborsky, L. Segments vs. whole sessions: Which is the better unit for psychotherapy process research? *Journal of Abnormal Psychology*, 1971, 78, 190-191.
- Mitchell, K. M., Bozarth, J. D., & Krafft, C. C. A reappraisal of the therapeutic effectiveness of accurate empathy, non-possessive warmth, and genuineness. In A. S. Gurman & A. M. Razin (Eds.), *Effective psychotherapy: A handbook*. New York: Pergamon Press, 1977.
- O'Leary, K. D., & Wilson, G. T. *Behavior therapy: Application and outcome*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.
- Paul, G. L. Behavior modification research: Design and tactics. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- Rathus, S. A. A 30-item schedule for assertive behavior. *Behavior Therapy*, 1973, 4, 398-406.
- Rice, L. N. Therapist's style of participation and case outcome. *Journal of Consulting Psychology*, 1965, 29, 155-160.
- Rich, A. R., & Schroeder, H. E. Research issues in assertiveness training. *Psychological Bulletin*, 1976, 83, 1081-1096.
- Robinson, J. P., & Shaver, P. R. *Measures of social psychological attitudes*. Ann Arbor, Mich.: Institute for Social Research, 1973.
- Ryan, V. L., & Gizynski, M. N. Behavior therapy in retrospect: Patients' feelings about their behavior therapies. *Journal of Consulting and Clinical Psychology*, 1971, 37, 1-9.
- Sloane, R. B., Staples, F. R., Cristol, A. H., Yorkston, N. J., & Whipple, K. *Psychotherapy versus behavior therapy*. Cambridge, Mass.: Harvard University Press, 1975.
- Truax, C. B., & Mitchell, K. M. Research on certain therapist interpersonal skills in relation to process and outcome. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis*. New York: Wiley, 1971.
- Watson, D., & Friend, R. Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, 1969, 33, 448-457.
- Watzlawick, P., Weakland, J., & Fisch, R. *Change*. New York: Norton, 1974.
- Wilson, G. T., & Evans, I. M. Adult behavior therapy and the therapist-client relationship. In C. M. Franks & G. T. Wilson (Eds.), *Annual review of behavior therapy. Theory and practice* (Vol. 4). New York: Brunner/Mazel, 1976.
- Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.
- Zuckerman, M., & Lubin, B. *Manual for the Multiple Affect Adjective Checklist*. San Diego, Calif.: Educational and Industrial Testing Service, 1965.

Received October 19, 1977 ■

Effects of Aging, Organicity, Alcoholism, and Functional Psychopathology on WAIS Subtest Profiles

John E. Overall, Norman G. Hoffmann, and Harvey Levin

Department of Psychiatry
University of Texas Medical Branch, Galveston

Multivariate analysis of variance was used to examine the independent effects of several factors on Wechsler Adult Intelligence Scale subtest patterns. A discriminant function contrasting organic brain syndromes with lesser functional psychiatric disorders was defined. Alcoholic patients were found to occupy a position close to the organic brain syndrome group on that continuum. Aging was associated with deficits in specific subtests, whereas the organic brain syndrome pattern involved more generalized deficits.

The primary purpose of this investigation was to examine the similarities and differences in Wechsler Adult Intelligence Scale (WAIS) profile patterns for various organic and functional psychiatric diagnostic groups, with the effects of several other relevant factors controlled statistically. In recent years, there has been increasing interest in the nature of changes in patterns of intellectual abilities associated with aging and with various organic and functional disorders, but the effects have seldom been clearly separated. Most authors agree that in cross-sectional studies at least, age decrements in intellectual performance begin to appear rather early (Jones, 1959; Wechsler, 1958). Such age-related changes tend to be confounded with the changes that are produced by disease states, which are also differentially age related.

The potential diagnostic contributions of the WAIS have been said to range from the differentiation between primary depression and the mood alterations that are frequently associated with senile dementia (Post, 1975) to the discrimination between patients with

Alzheimer's disease and those having a strictly functional psychiatric disorder (Malamud, 1975). The possible differential diagnostic significance of the WAIS in elderly patients is supported by the findings of a number of other investigators who have correlated changes in WAIS profile patterns with estimated degree of cerebral integrity (Thompson, 1976), slowing of dominant frequency of the electroencephalogram (Wang, 1973), and with reduced cerebral blood flow (Wang, Obrist & Busse, 1970). Most studies suggest that the Performance subtests that correspond to the factor of "perceptual organization" (Cohen, 1957) are particularly sensitive to neurologically based intellectual deterioration in elderly patients. A problem with interpretation of such results is that aging, from very early on, produces pattern changes in the absence of any evidence of neuropathology. In most studies, the effects that have been reported in elderly patients suffering from neurological disfunctions of one type or another are almost certainly confounded with the effects of normal aging, even though age-group norms have frequently been considered.

Somewhat different from the question of diagnostic utility is the hope that subtle differences in profile patterns can lead to a better understanding of the nature of cognitive, perceptual, and phenomenological deficits experienced by psychiatric patients. Overall and Gorham (1972) questioned

This work was supported in part by U.S. Department of Health, Education, and Welfare Grant DHEW 5 R01 MH14675-10.

Requests for reprints should be sent to John E. Overall, Department of Psychiatry, University of Texas Medical School P.O. Box 20708, Houston, Texas 77025.

whether the impairment present in normally aging individuals is similar to that observed in patients with diagnoses of chronic brain syndrome. Using the method of multiple discriminant analysis with WAIS subtest profiles, they found that normally aging individuals seemed to differ on one dimension, whereas the chronic brain syndrome patients differed from the normal aging groups along a separate dimension. Specifically, in the institutionalized samples that they studied, simple aging produced greater pattern variability, whereas old patients with chronic brain syndromes evidenced a more general deficit that was reflected in both Verbal and Performance subtests.

Subsequently, Williams, Ray, and Overall (1973) used the WAIS aging and organicity functions defined in Overall and Gorham (1972) to investigate the nature of possible subtle changes in the phenomenological world of the alcoholic patient. They questioned whether the preclinical changes associated with alcohol abuse are more consistent with accelerated mental aging or with a developing organic brain syndrome. They found that the WAIS subtest profiles for alcoholics in a state hospital sample deviated from age norms along both the mental aging and organicity dimensions. Such findings are compatible with other work involving the effects of chronic alcohol abuse in intellectual functioning (Matarazzo, 1972).

The present study attempts to use a more appropriate and powerful multivariate statis-

tical method to separate out and clarify the nature of WAIS profile pattern changes that are associated with chronic organic brain syndrome, alcoholism, and various functional psychiatric conditions independent of the effects associated with aging and with background factors that might also contribute to apparent diagnostic group differences. The need for such complex statistical methods arises because diagnostic groups differ in average age and in other factors that are known to affect WAIS profile patterns. Left uncontrolled, such factors can produce differences in the WAIS profile patterns that are in no way characteristic of the disease processes themselves.

A multivariate analysis of variance (MANOVA) procedure was used to partial out the effects of several partially confounded nuisance variables, to estimate the WAIS profile patterns typical of diagnostic groups with age held constant, and to estimate the effects of simple aging with diagnostic group differences held constant. The prototype for the MANOVA is a factorial design with equal numbers of subjects in all cells. The equal cell frequencies ensure that the effects of any particular factor, such as age, are equally represented in each diagnostic group. In practice, it is not necessary to have each factor balanced with regard to presence in the various levels of each other factor. A nonorthogonal design can be analyzed by least squares regression methods to obtain estimates of the same effects that would be observed in a balanced (orthogonal) design involving the same factors (Overall, Spiegel, & Cohen, 1975).

Table 1
Twelve Major Diagnostic Groups and Their Frequencies

Diagnosis	n
Organic brain syndrome	29
Drug abuse	13
Alcoholic	29
Paranoid schizophrenia	43
Schizophrenia	52
Manic	20
Schizoaffective—manic	16
Schizoaffective—depressed	41
Psychotic depression	29
Depressive reaction	91
Situational reaction	33
Personality disorder	18

Method

WAIS subtest profiles for a sample of 414 psychiatric patients were obtained from the files of the Psychometric Laboratory at the University of Texas Medical Branch. Final clinical diagnoses were used to group the patients into 12 major diagnostic categories, as listed in Table 1. The organic group is of special interest in this investigation, since it represents a generally younger and somewhat less chronic sample with regard to the question of differences between mental aging and organicity than previously studied by Overall and Gorham (1972). The organic brain syndrome group in the present case includes patients who were given diagnoses of either

Table 2
Summary of Tests for Main Effects in a Five-Way MANOVA Design

Variable	Hotelling trace		Wilk's Λ	Pillai-Bartlett	
	χ^2	df		F	df
Age	142.16	33	.6942	4.19**	33, 1158
Ethnicity	118.37	22	.7376	5.56**	22, 770
Sex	84.50	11	.8045	8.48**	11, 384
Parental SES	39.86	22	.9026	1.80*	22, 770
Diagnosis	204.25	121	.5951	2.33**	121, 4334

Note. MANOVA = multivariate analysis of variance; SES = socioeconomic status.

* $p < .05$.

** $p < .001$.

chronic or acute organic brain syndrome, but all were either newly admitted inpatients or outpatients who were referred for diagnostic testing. No focal lesions, trauma, or temporal lobe epilepsy diagnoses were included.

The independent effects of age and diagnosis were of primary interest in the analyses of the WAIS subtest profiles for these psychiatric patients. Sex and race of the patient and social class of the patients' parents were selected as control variables that should substantially affect the expected IQ and profile pattern without being influenced by the cognitive abilities or clinical status of the patient. The desire was to control statistically for premorbid differences so that the actual effects of age and psychopathology could be seen more clearly without partialing out anything that might represent an effect of the variables of interest. A five-way Age \times Race \times Sex \times Socioeconomic Status (SES) \times Diagnosis MANOVA was used to test the effects of each factor independent of effects on WAIS profile patterns that might be attributed to other factors in the design. Univariate analyses of variance were also

calculated for each WAIS subtest using the same five-way design. A computer program by Woodward and Overall (1974) was used to accomplish the multivariate and univariate analyses in a single pass.

Results

The MANOVA resulted in the recognition that age, race, sex, and SES all have highly significant independent effects on WAIS subtest profiles and that WAIS subtest profiles for the 12 clinical diagnostic groups differ significantly after the effects of the other factors are partialled out. A summary of results from the tests of main effects in the five-way MANOVA is presented in Table 2. Three different test statistics are shown for each effect: an approximate chi-square test based on the sum of all roots of the $|B - \lambda W|$

Table 3
Adjusted Mean WAIS Subtest Profiles for Four Age Groups
in the Psychiatric Population

Subtest	<30	30-39	40-49	50+	F ^a
Information	7.84	8.22	7.84	7.98	.65
Comprehension	8.01	8.71	7.85	7.49	2.43
Arithmetic	7.07	7.48	7.03	6.61	1.31
Similarities	9.50	8.62	7.77	7.20	9.91*
Digit Span	7.72	8.36	7.44	7.25	2.35
Vocabulary	8.32	8.54	7.86	8.03	.97
Digit Symbol	7.80	6.97	5.56	4.47	28.11*
Picture Completion	8.41	8.20	7.00	6.36	11.46*
Block Design	7.69	7.48	6.23	5.58	9.93*
Picture Arrangement	7.95	7.75	6.54	6.19	7.02*
Object Assembly	7.90	7.63	6.68	6.05	5.66*

Note. WAIS = Wechsler Adult Intelligence Test.

^a df = 3, 394.

* $p < .01$.

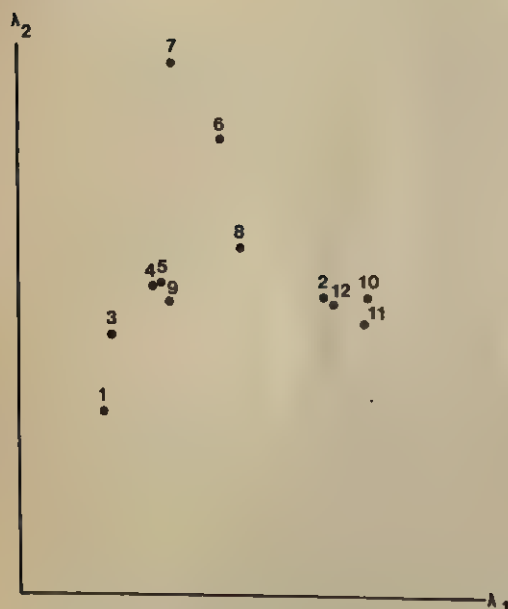


Figure 1. Configural relationships among diagnostic groups (organic brain syndrome, 1; drug abuse, 2; alcoholic, 3; paranoid schizophrenia, 4; schizophrenia, 5; manic, 6; schizoaffective—manic, 7; schizoaffective—depressed, 8; psychotic depression, 9; depressive reaction, 10; situational reaction, 11; and personality disorder, 12.)

$= 0$ determinantal equation (Hotelling, 1951), a chi-square test based on Wilks' likelihood ratio criterion lambda (Wilks, 1932), and the Pillai-Bartlett F approximation (Olson, 1976). The latter test is widely accepted as best for MANOVA purposes, although in the present case all tests lead to the conclusion that each factor has a highly significant independent effect on WAIS subtest profiles.

Adjusted mean subtest profiles for patients in the four age groups are presented in Table 3, together with the univariate F ratios for testing the significance of effects of age on each subtest separately. The interpretation of "adjusted means" in a nonorthogonal design should perhaps be clarified at the outset. The adjusted means are estimates of the means that would obtain if the total sample included equal numbers of males and females, as well as equal numbers in each of the three social classes, the three ethnic groups, and the 12 diagnostic groups. They statistically estimate the same effects that one might otherwise estimate by sampling in such a

manner that all cells in the complex factorial design contained the same N (Overall et al., 1975). It is in this manner that the effects of other factors are balanced out so that the effect of any one factor can be evaluated independently of any confounding with other factors in the design.

It is apparent in Table 3 that age had selective effects on WAIS subtest profiles in the psychiatric population, just as has been reported in the general population. The age effects were pronounced on all Performance tests and on the Similarities subtest of the Verbal section. Age effects were not observed on Information, Arithmetic, and Vocabulary subtests, and there was only a nonsignificant trend toward a decrement in Comprehension and Digit Span. Thus, considering that the Similarities subtest is the only Verbal subtest requiring "fluid intelligence" (Cattell, 1963), deficits associated with aging in the psychiatric population are clearly represented in an increase in the discrepancy between "crystallized" verbal abilities and cognitive-perceptual-motor behavior requiring flexibility, speed, and agility.

Adjusted subtest means for the 12 clinical diagnostic groups are presented in Table 4. In this case, the profiles have been adjusted to reflect a common average age, race, sex, and SES. Independent of age and the other factors, patients in the different diagnostic groups evidenced significantly different WAIS profile patterns. The diagnostic group differences can be seen to be represented in both Verbal and Performance subtests to a greater degree than was true with the age effects. Again, the diagnostic effects represented in these adjusted profiles are associated with psychopathology *independent* of age and other factors. Whereas age effects were absent in the Information, Arithmetic, and Vocabulary subtests, significant diagnostic group differences were present on those subtests, as well as on the Performance subtests, which were also sensitive to age. No significant diagnostic group differences were present for Similarities, which was the only Verbal subtest that was affected by age.

The MANOVA between diagnostic groups is equivalent to a multiple discriminant function analysis calculated on WAIS profiles from

which the effects of age, race, sex and SES have been eliminated. The analysis confirmed that all of the significant group differences were represented in two major dimensions. The sum of all roots of the error-adjusted between-groups matrix was $\Sigma\lambda_i = 217.2$, which can be referred to a chi-square distribution with $p(k-1) = 121$ degrees of freedom. The first and second roots were $\lambda_1 = 79.8$ and $\lambda_2 = 41.4$, with $(p+k-2)$ and $(p+k-4)$ degrees of freedom, respectively. The sum of all remaining roots, after the first two were subtracted out, was $\Sigma\lambda_i = 96.0$, which is approximately distributed as chi-square with 81 degrees of freedom, and was not statistically significant. Thus, the data provide no evidence that more than two dimensions are required.

The configuration of group means in the plane defined by the first two discriminant functions is displayed graphically in Figure 1. The first, or horizontal, dimension contrasts the organic group with several of the least impaired of the functional psychiatric groups. It is defined primarily as a contrast between Information and Vocabulary on the one hand and Arithmetic and Digit Symbol on the other. Individuals who score toward the organic end of this contrast function tend to score low on Arithmetic and Digit Symbol *relative* to their scores on Information and Vocabulary. This interpretation can be verified by calculating the simple contrast from the adjusted group means in Table 4.

The second, or vertical, dimension separates the organic group from the more accelerated schizoaffective, manic, and manic-depressive groups. It tends to emphasize a general deficit in performance of individuals who score toward the organic end of the vertical continuum. If one considers the *sum* of Vocabulary and Digit Symbol scores, rather than their difference, the ordering of diagnostic groups in the vertical dimension can be reasonably well approximated. The ordering in the vertical dimension can be even more closely approximated by the sum of adjusted mean scores on all 11 subtests in Table 4.

In considering the question of a general deficit, as well as the more specific pattern effects, it is important to recognize that the

Table 4
Adjusted Mean WAIS Subtest Profiles for 12 Diagnostic Groups in the Psychiatric Population

Subtest	Organic brain syn- drome	Drug abuse	Alco- holism	Paranoid schizo- phrenia	Schizo- phrenia (non- paranoid)	Manic	Schizo- affective manic	Schizo- affective depressed	Psy- chotic depres- sion	Depres- sive reaction	Situa- tional reaction	Person- ality disorder	F
Information	7.88	8.22	8.29	6.93	7.80	8.48	9.48	8.34	7.47	7.85	7.38	7.52	1.95*
Comprehension	8.73	8.68	8.05	6.94	7.04	8.16	9.18	7.95	7.47	8.10	7.38	8.52	1.71
Arithmetic	6.78	8.28	6.47	6.24	6.58	6.85	7.16	7.33	6.43	7.73	7.17	7.56	1.83*
Similarities	7.63	8.83	8.19	7.59	7.59	8.70	9.32	8.35	7.56	8.52	8.13	8.85	1.21
Digit Span	6.94	7.10	6.70	7.27	7.44	7.86	8.62	8.16	7.56	8.25	7.95	8.45	1.41
Vocabulary	7.77	8.01	8.61	7.69	7.65	9.14	10.25	8.32	7.75	7.89	7.25	7.91	2.12*
Digit symbol	4.58	6.98	5.26	5.29	5.87	6.97	7.12	6.15	5.43	6.89	6.83	7.05	4.14**
Picture Completion	7.02	7.39	7.96	6.50	6.37	7.63	8.12	7.71	6.94	7.93	8.08	8.30	3.05**
Block Design	6.18	6.60	5.85	6.34	6.17	6.65	8.40	7.20	6.17	7.03	6.31	8.04	2.18*
Picture Arrangement	5.55	7.73	6.61	6.58	6.61	6.93	8.18	7.76	6.60	7.73	7.34	7.68	2.28*
Object Assembly	6.22	7.53	5.74	6.26	5.60	8.25	8.57	7.84	6.45	7.30	7.50	7.47	3.34**

Note. WAIS = Wechsler Adult Intelligence Scale. * $p < .05$. ** $p < .01$.

WAIS subtest profiles in this analysis were adjusted statistically for the effects of age, race, sex, and parental social class. It is difficult clinically to take all relevant factors into consideration in deciding whether a general or specific pattern deficit is present. The aim here is merely to understand what the effects of organic and functional psychopathology are on the cognitive-perceptual-motor performance of psychiatric patients. The results suggest that organic patients manifest a general impairment, which is most marked on tasks requiring perceptual-motor speed and agility. As compared with other types of psychopathology, moderate hypomania appears to have a less detrimental effect on WAIS performance.

It is interesting to note the proximity of the alcoholic group to the clinically organic group in the WAIS subtest space. None of the alcoholic patients had a clinically recognizable organic brain syndrome; yet, as a group they ranked next to the organic brain syndromes on both the specific (horizontal) and general (vertical) dimensions of deficit. Again, it would be difficult to evaluate such trends clinically because of the need to control for other factors that influence WAIS performance; however, the results do suggest that WAIS profile patterns are sensitive to subtle pre-clinical changes associated with alcohol abuse and that the brain syndromes that occur as end points in alcoholism are not precipitous outcomes.

Ethnicity and sex of patients and social class of patients' parents were included primarily as control variables to adjust statistically for major premorbid differences that should not be confused with the effects of age or psychopathology. As a consequence, the results pertaining to the effects of those factors will be discussed only very briefly. It should be noted that educational achievement was not partialled out because it is in part a result of IQ.

The total patient sample was subclassified as white, black, and Mexican American. Two significant dimensions were found to separate the three groups in the WAIS subtest space after the effects associated with age, sex, parental SES, and clinical diagnosis were partialled out. The first major dimension sepa-

rated the three groups in the order white, Mexican American, and black. It was a general factor that correlated highly with the total score on all subtests. The second discriminant function separated Mexican Americans from the other two groups, and it tended to be a contrast between Verbal and Performance subtests, with the Mexican Americans scoring relatively better on the Performance subtests than on the Verbal subtests.

Sex differences were apparent in this psychiatric population after adjustment was made for age, race, parental SES, and clinical diagnosis. Males scored significantly higher on Information, Arithmetic, and on all Performance subtests except Digit Symbol. Females did not score significantly higher on any subtest.

Social class of parental home was taken from a history form completed by the psychometrists who did the testing. Although it was routinely inquired about, there is some concern that inferences based on social achievement of the patients themselves may have crept into the scoring of this item. Social class was scored as lower, lower middle, or middle, and upper. After adjustment was made for age, race, and diagnosis, the effect of the SES factor was highly significant on all subtests. A single significant dimension separated the groups in the anticipated order, with Information and Vocabulary being the two most discriminating variables.

The multiway MANOVA provided evidence that diagnostic groups differ in WAIS profiles independently of age and that age groups differ independently of diagnosis. It did not, however, provide a basis for concluding whether the *pattern* of effects produced by aging is different from the pattern associated with organic brain syndrome. Neither of the two discriminant dimensions in Figure 1 can be considered to be *the* organicity dimension, since the organic brain syndrome group differed from the other groups along both dimensions. A 45-degree rotation of the horizontal axis in Figure 1 was used to define a function that tends to separate the organic group from all other groups along a single dimension in the age-corrected WAIS subtest space. The MANOVA directly provided a single age function in the diagnosis-corrected WAIS

subtest space. The weighting coefficients defining these two discriminant functions are presented in Table 5.

Further attention can now be directed to whether mental aging and organicity involve similar or different profile pattern effects in the psychiatric population. It has already been noted in Table 4 that diagnostic groups differed significantly on certain Verbal subtests that were not significantly affected by age as shown in Table 3. Examination of the weighting coefficients in Table 5 reveals that the aging function is more clearly a contrast between Verbal and Performance subtest scores, whereas the organicity function contrasts subtests within the two major domains. The largest effect of both aging and organicity was on the Digit Symbol subtest; however, low Vocabulary and Digit Span scores were positive indicators for organicity but negative indicators for mental aging independent of diagnosis. The variability in profile pattern, and therefore the variability of weighting coefficients in the discriminant function, is greater for mental aging effects than for the effects of organic brain syndrome.

From a conceptual/interpretative point of view, the results can be understood by considering the age effect to increase *pattern variability* through reduction in "fluid intelligence" scores and the organic brain syndrome effect to represent a more general *global deficit*. The age groups order in a similar manner in both dimensions because the "fluid intelligence" deficit, which increases pattern variability, also reduces total scores. The organic brain syndrome effect is almost entirely in the dimension characterized by overall deficit and only slightly in the direction of increased pattern variability.

Discussion

The present study supports earlier work, which indicated that the effects of organic involvement on the cognitive, perceptual, and motor functions in patients diagnosed as having organic brain syndrome are distinct from the effects of the normal aging process.

Differences attributable to aging in this psychiatric sample appeared to fit well into the crystalized versus fluid intelligence con-

Table 5
Discriminant Function Coefficients Defining Independent Mental Aging and Organicity Effects in WAIS Subtest Profiles

Subtest	Aging	Organicity
Information	.1652	.0693
Comprehension	.0663	.0830
Arithmetic	.1141	.0326
Similarities	-.1936	-.0331
Digit Span	.0453	-.0687
Vocabulary	.0499	-.0566
Digit Symbol	-.3399	-.3069
Picture Completion	-.1035	.0626
Block Design	-.0400	.0660
Picture Arrangement	-.0526	-.1237
Object Assembly	.0005	-.1743

Note. WAIS = Wechsler Adult Intelligence Scale.

cept proposed by Cattell (1963), in that older patients tended to show lower scores on those subtests that required mental flexibility and perceptual-motor speed relative to those subtests that tap established or practiced cognitive abilities. In contrast, organic patients manifested a more general deficit in intellectual functioning as measured by the WAIS, as well as a different pattern of relative score deficits.

The thrust of this study has been the investigation of WAIS subtest scores in an attempt to provide some understanding of the nature of defects found in psychiatric patients. As such, it was not intended to directly provide aspects of clinical utility or to support specific clinical approaches. The statistical adjustments made in these analyses are not readily applicable for clinical use. Also, significant results in group data do not necessarily imply sufficient discriminative power for application to individual cases. Yet, the results reported here would appear to have some implications for further work along both theoretical and clinical lines.

One such approach might be the replication of this study on a sample of patients with selected organic and functional disorders in conjunction with more neuropsychological data on the patients. This would facilitate the consideration of specific types of organic involvement and the assessment of apparent cognitive deficits in specific diagnostic groups. As was noted in this study, patients with the

diagnosis of alcoholism appeared most similar to the organic group, even though alcoholics with any definite or possible organic diagnoses were excluded from the sample. Similarly, schizophrenics were also found to be more similar to the organic group than were those with affective psychoses and the less disturbed psychiatric patients. This would perhaps suggest confirmation of the work of Johnstone, Crow, Frith, Husband, and Kreel (1976), who found evidence of enlarged ventricles in the in vivo brains of schizophrenic patients.

The current results showing distinct effects of age support the use of age-scaled standard scores as provided in the WAIS manual or other standardized scores adjusted for age when the WAIS is to be used in the assessment of organicity.

One can also infer differential utility of the 11 subtests in the evaluation of organicity. Some of the WAIS subtests seem more effective in tapping those abilities or capabilities that are diminished by organic impairment. Perhaps a few subtests could be used to answer specific questions relative to specific types of organic involvement. On the other hand, further work might indicate that the WAIS is not readily adaptable to such a task, but other tests or instruments that provide better measurement of the subtle distinctions suggested here can be devised to determine the degree of organic impairment.

Research containing multivariate techniques such as those used in this study may provide increased understanding of cognitive, perceptual, and motor deficits in psychiatric patients. Some efforts may be primarily of theoretical import, but others could make contributions to clinical practices.

References

- Cattell, R. B. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 1963, 54, 1-22.
- Cohen, J. The factorial structure of the WAIS between early adulthood and old age. *Journal of Consulting Psychology*, 1957, 21, 283-290.
- Hotelling, H. A generalized *T* test and measure of multivariate dispersion. *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. Berkeley: University of California Press, 1951.
- Johnstone, E. C., Crow, T. J., Frith, C. D., Husband, J., & Kreel, L. Cerebral ventricular size and cognitive impairment in chronic schizophrenia. *Lancet*, 1976, 2, 924-926.
- Jones, H. E. Intelligence and problem solving. In J. E. Birren (Ed.), *Handbook of aging and the individual*. Chicago: University of Chicago Press, 1959.
- Malamud, N. Organic brain disease mistaken for psychiatric disorder: A clinicopathologic study. In B. F. Benson & D. Blumner (Eds.), *Psychiatric aspects of neurologic disease*. New York: Grune & Stratton, 1975.
- Matarazzo, J. D. *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). New York: Williams & Wilkins, 1972.
- Olson, C. L. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 1976, 83, 579-586.
- Overall, J. E., & Gorham, D. R. Organicity versus old age in objective and projective test performance. *Journal of Consulting and Clinical Psychology*, 1972, 39, 98-105.
- Overall, J. E., Spiegel, D. K., & Cohen, J. Equivalence of orthogonal and nonorthogonal analysis of variance. *Psychological Bulletin*, 1975, 82, 182-186.
- Post, F. Dementia, depression, and pseudo-dementia. In D. F. Benson & D. Blumner (Eds.), *Psychiatric aspects of neurologic disease*. New York: Grune & Stratton, 1975.
- Thompson, L. W. Cerebral blood flow, EEG, and behavior in aging. In R. D. Terry & S. Gershon (Eds.), *Aging*. New York: Raven Press, 1976.
- Wang, H. S. Cerebral correlates of intellectual function in senescence. In L. F. Jarvik, C. Eisdorfer, & V. E. Blum (Eds.), *Intellectual functioning in adults*. New York: Springer, 1973.
- Wang, H. S., Obrist, W. D., & Busse, E. W. Neurophysiological correlates of the intellectual function of elderly persons living in the community. *American Journal of Psychiatry*, 1970, 126, 1205-1212.
- Wechsler, D. *The measurement appraisal of adult intelligence*. (4th ed.). Baltimore: Williams & Wilkins, 1958.
- Wilks, S. S. Certain generalizations in the analysis of variance. *Biometrika*, 1932, 24, 109-116.
- Williams, J. D., Ray, C. G., & Overall, J. E. Mental aging and organicity in an alcoholic population. *Journal of Consulting and Clinical Psychology*, 1973, 41, 392-396.
- Woodward, J. A., & Overall, J. E. A general multivariate analysis of variance computer program. *Educational and Psychological Measurement*, 1974, 34, 653-662.

Locus of Control in Mexicans and Chicanos: The Case of the Missing Fatalist

David Cole and Jacqueline Rodriguez
Occidental College

Shirley Cole
University of Arizona

We studied the extent to which a stereotype of Mexican or Chicano students as fatalistic is supported by their locus of control scores. Data came from locus of control scores from male university students in four nations: United States (86), Mexico (57), Ireland (47), and West Germany (54). These data show the Mexican university students to be more internally oriented than students from each of the other nations ($p < .001$). The study also compares locus of control scores for 151 Anglo and 95 Chicano senior high school students from three Southern California high schools. Scores for Chicanos are nearly identical to those obtained from Anglo students. Only Chicano male high school students not planning to enter college showed any tendency toward a more external locus of control ($p < .05$). The article concludes that to the extent a perceived external locus of control would be indicative of a fatalistic outlook, such perception is lacking in most data on Mexican and Chicano respondents.

Recent years have seen several studies allowing transnational comparisons, usually among students, in locus of control measures (Hsieh, Shybut, & Lotsof, 1969; Mahler, 1974; McGinnies, Nordholm, Ward, & Bhanthumnavin, 1974; Parsons & Schneider, 1974; Schneider & Parsons, 1970; Reitz & Groff, 1974). Resulting differences between persons of different nations are then related to different societal structures, socialization patterns, or national traditions or customs. Thus, for example, Hsieh et al. (1969) ascribed the greater externality found among their Chinese respondents, when contrasted to American-born Chinese or Anglo Americans, to differences in outlook that are traditional to Chinese as contrasted with Americans. McGinnies et al. (1974) provided a series of suggestions based on Japanese culture and society for their finding of greater externality

among Japanese respondents as contrasted to those from three nations with an Anglo-Saxon tradition. They suggested that the Swedish social system of maximizing personal security may account for the relatively external scores of their Swedish respondents, but they also noted that the Swedish respondents were younger than those from other national groups in their study.

Among the studies cited above, only that by Reitz and Groff (1974) used other than student respondents. Reitz and Groff focused on locus of control as a function of the degree of economic and industrial development in a nation and used nonsupervisory laborers from the United States, Mexico, Japan, and Thailand as respondents. There were no significant main effects of the state of economic development, although differences did appear on the subscales of the Rotter test (Rotter, 1966). There was a significant main effect when the western nations were compared to the Asian nations, with the western nations being overall more internally oriented. Comparing the United States with Mexico, the Americans were more

This study was adapted from a paper presented at the Rocky Mountain Psychological Association, Phoenix, Arizona, May 1976.

Requests for reprints should be sent to David Cole, Department of Psychology, Occidental College, Los Angeles, California 90041.

Table 1
Mean Locus of Control (I-E) Scores

Group	n	M
U.S. business administration	38	10.03
U.S. liberal arts	48	10.15
Germany	54	9.24
Ireland	47	10.23
Mexico	57	5.88

Note. $P = .001, p < .001$

internal on the Leadership subscale and more external on the Luck subscale, with no overall significant differences between persons from the two nations.

The work of Reitz and Groff (1974) is particularly relevant to the present article, because it included persons from Mexico and allowed evaluation of the extent to which such persons may be described as "fatalistic" when compared to respondents from other nations. The present study, which began before the Reitz and Groff (1974) data appeared, is an attempt to evaluate the validity of the stereotype of fatalistic as it applies to Mexican and Chicano students. We were initially moved to a study of this question by two related observations. When discussing the concept of locus of control as it might appear in Mexican and/or Chicano respondents, Anglo friends of ours were usually quick to predict that such persons would score in an *external* direction as contrasted to Anglo Americans, whereas the Chicano who we consulted usually declared that this would be the case.

The importance of this issue should not be understated. Numerous authors (e.g., Berry, 1964; Coon & Makin, 1971; D. Anderson, 1975; Miller, 1966; Justin, 1970; Lewis, 1970; Morales, 1971) have shown that Mexican and/or Mexican Americans who are viewed by Anglos as passive and fatalistic. In general, these authors have not provided extensive empirical support for the stereotype. For example, in their studies of students in U.S. universities, these authors have concentrated on the stereotype that Anglo Americans have of Mexico, Jews or persons from various European countries (e.g., Miller, 1966; Reitz, 1974; & Watters, 1970; Katz & Brady, 1970). To determine if such a stereo-

type exists among American students similar to those studied in this investigation, a sample of 53 students at the liberal arts college (cited in Table 1) were asked to complete the Rotter scale as they believed they would answer if they were Mexican students attending a Mexican university. In contrast to the mean of 5.88 actually obtained from Mexican students, or the mean of 10.15 obtained from American liberal arts students answering for themselves, these "pseudo-Mexican" students produced a mean of 15.55. This is significantly more *external* than the mean of actual Mexicans ($p < .001$) or Americans answering for themselves ($p < .01$). Combining these data with the observations of the several authors cited above, we contend that the presence of a stereotype of Mexicans and/or Mexican Americans as passive and fatalistic is justified. In many geographic areas of the United States, and especially those from which we come, counselor and teacher interaction with Chicano and Mexican students is widespread and daily. There is ample reason to investigate the validity of a stereotype that is likely to influence counselor and teacher prediction of a student's academic behavior.

The contention that perception of an external locus of control would indicate a fatalistic outlook seems warranted by what is implied by such a perception. If one believes, as a person with a perceived external locus of control is presumed to believe, that positive and/or negative events in one's life are beyond one's personal control, this seems clearly to invite, if not define, a fatalistic outlook.

Method

Subjects

For the study reported here, the Rotter Internal-External Locus of Control (I-E) Scale (Rotter, 1966) was administered to male, Catholic, business administration students in the United States, West Germany, Ireland, and Mexico. For use in West Germany and Mexico, the scale was translated into the appropriate language, following procedures recommended by Brulin (1970). The Rotter scale was

The authors wish to express appreciation to Robert Chestnut who arranged for the translation of the scales into German.

first translated into the appropriate foreign language (Spanish or German) by a bilingual psychologist, whose first language is English. The translation was then given to a bilingual colleague, whose first language is Spanish or German, respectively. This colleague translated the scale back into English, without seeing the original scale. Translation was considered accurate when this translation back into English corresponded with the original English version. When results from Mexico proved striking, the translation process was repeated, as a double check on possible errors. No changes were made as a result of this duplication of the initial work. By selecting business administration majors, we obtained students in the four nations who were pursuing highly similar curricula, with similar career goals. In limiting the respondents to those who were at least nominally Catholic, we attempted to control for the potentially relevant variable of religious frame of reference as a factor influencing one's perceived locus of control. When data analysis showed these two variables to be irrelevant, we added a second group of U.S. students, from a liberal arts college, to expand the base of the American sample, but we kept this second group separated from the American business administration students throughout the data analysis. Because sex has been shown to be a variable in I-E scores (McGinnies et al., 1974) and because no females were available from Ireland and very few from West Germany, we confined the sample to males. All scales were administered in the classroom, by native speakers. The Irish sample came from a university in Dublin, the German from a university in the western portion of West Germany, and the Mexican sample from a university in south-central Mexico. The American business administration students came from a private, Catholic university on the west coast, and the liberal arts students came from a private liberal arts college also on the west coast. West German and Irish students were included because they were available to us, and they provided, along with the American data, a much broader base of national groups against which to compare and contrast the Mexican data.

At the outset of the study, the Levenson scale (Levenson, 1974) was not available. It was available by the time we started gathering data from Chicanos, however, and its ready analysis into three separate scales made it worthwhile to use, particularly since direct score comparison with the university student groups was not needed. For the high school portion of the study, the Levenson scale was administered in classrooms to 12th-grade students in three Los Angeles area high schools, one almost entirely Anglo and two almost entirely Chicano in makeup of student population. In the Chicano high schools, three different schedules of administration were used: In some classes a Chicano psychologist administered the scales, in some an Anglo psychologist did so, and in some both administrators were present. Ethnicity of the scale administrator proved irrelevant. All scales in the

Anglo classrooms were administered by an Anglo. In all three schools, data were collected from students in college preparatory curricula and from students not in such curricula.

Results

The results from the testing of university students in the four nations are presented in Table 1.

The mean score for the Mexican respondents was significantly lower than each of the other means ($p < .001$, in each instance). No significant differences were found between the means of any of the other national groups.

Because Reitz and Groff (1974) presented their data in terms of the percentage of externally oriented answers given on each of the five Rotter subscales developed by Schneider and Parsons (1970), we made the same breakdown in analyzing the present data. Our figures are presented in Table 2. Tests for significant differences in proportions were tested by z scores, again following the method of Reitz and Groff. Beyond those cited in Table 2, only one other comparison reached significance. The U.S. liberal arts students scored more external on the Respect subscale than did the West Germans. Except for this, all significant differences attest to the greater internality of the Mexican respondents.

Table 3 presents the data on Mexican laborers as reported by Reitz and Groff (1974), alongside our own data from university students in the United States and Mexico. On each subscale the Mexican students were more internal than the Mexican laborers. Both groups of American students were more internal than American laborers on the Politics subscale and more external on the Respect subscale. One group of American laborers was more external than American laborers on the Leadership subscale, but this did not hold for the other student group.

The Levenson scale provides three scores: internal locus of control, control by powerful others, and control by chance. Scores on each scale can range from -24 to $+24$. Results from the testing of high school students are presented in Table 4.

There are no significant differences, on any of the scales, between Chicano and Anglo

Table 2

Comparisons of Percentage of External Responses by Category

Group	Category				
	Luck	Politics	Respect	Academic	Leadership
	<i>M % external</i>				
U.S. business administration	41	44	57	49	30
U.S. liberal arts	40	40	62	46	40
Germany	38	39	39	45	43
Ireland	45	54	46	28	43
Mexico	19	38	26	20	26
	<i>z scores of differences in % external</i>				
Mexico: U.S. business administration	2.44**	<i>ns</i>	3.10***	3.02***	<i>ns</i>
Mexico: U.S. liberal arts	2.50**	<i>ns</i>	3.78***	2.88***	<i>ns</i>
Mexico: Germany	2.23*	<i>ns</i>	<i>ns</i>	2.87***	2.27*
Mexico: Ireland	2.88***	<i>ns</i>	2.15*	<i>ns</i>	2.17*

Note. Only one other comparison reached significance. U.S. liberal arts students scored in a more external direction on the Respect scale than Germans ($z = 2.34$, $p < .05$).

* $p < .05$.

** $p < .02$.

*** $p < .01$.

students. Within ethnic groups, the college-bound Chicano male rejected control by chance more than did his non-college-bound counterpart ($p < .05$). No significant sex differences were found. All of the high school groups, whether Anglo or Chicano, tended to endorse the idea of internal locus of control and to reject the ideas of control by powerful others or chance happenings.

Discussion

This study began as an attempt to evaluate the validity of the stereotype of Mexicans and Chicanos as fatalistic, as that attitude would be expressed through perception of an external locus of control. Clearly, the Mexican university students were *not* more external; indeed, they were significantly more

Table 3

Comparison of Students and Factory Workers in the United States and Mexico

Group	Category			
	Luck	Politics	Respect	Leadership
	<i>M% external</i>			
Mexican factory workers	34	56	38	47
Mexican students	19	38	26	26
U.S. factory workers	42	58	43	30
U.S. business administration students	41	44	57	30
U.S. liberal arts students	40	40	62	40
	<i>z scores of differences in % external</i>			
Mexican students: Mexican factory workers	4.50*	5.05*	5.05*	6.77*
U.S. factory: business administration	.34	4.79*	4.77*	0
U.S. factory: liberal arts	.62	5.53*	5.81*	3.31*

Note. Factory worker data are taken from Reitz and Groff (1974), who did not report data on the Academic scale.

* $p < .001$.

Table 4
Locus of Control Scores: High School Seniors

Group	n	College			n	Noncollege		
		Internal	Powerful others	Chance		Internal	Powerful others	Chance
Males								
Anglo	47	9.86	-3.88	-5.68	28	11.29	-1.11	-3.46
Chicano	23	11.00	-3.74	-5.74*	24	11.04	-1.86	-1.29*
Females								
Anglo	38	10.52	-3.24	-5.94	38	10.11	-3.29	-3.71
Chicana	29	11.41	-3.24	-3.52	29	11.33	-4.58	-3.52

* $p < .05$.

internal in locus of control than students from three other nations. The finding is not limited to a single time period, for the data were collected at two points in time, about 7 months apart. The results are not limited to the particular instrument, for two of us (Cole & Cole, 1977) have reported a similarly strong emphasis on internal locus of control using the Levenson scale with Mexican students. The data from Reitz and Groff (1974) allow us to note that this is not an artifact of testing students. Their Mexican laborers, although not more internal than American laborers, were not more external either. Indeed, on the Luck subscale, surely the psychological heart of fatalism, both Mexican students and Mexican laborers received their least externally oriented scores. Thus we feel that these data lend no support to, and instead challenge, the stereotype of Mexicans as fatalistic.

Turning to Chicanos, we know of only one previous study that addressed this issue. Garza and Ames (1974) reported that Chicano university students scored in a significantly more internal direction than Anglo students in the same university. Our data for high school students do not show a group difference, but they are striking in the high degree of similarity between the scores from Anglos and from Chicanos. Only the male Chicanos not planning to enter college offered the most tenuous support for a fatalistic stereotype, in their significantly lower rejection of control by chance. Even here however, it is important to note that they did

not embrace chance; they merely rejected it less than other students.

The findings regarding Mexicans and Chicanos seem so at odds with Anglo folk wisdom that a further item analysis seemed indicated. Gurin, Gurin, Lao, and Beattie (1969) have distinguished between perceived locus of control as it applies directly to the self and as a perception of a general, but not necessarily personal, social condition. Consequently, we did an item analysis that contrasted responses to the personally worded items of the Rotter scale with those describing general social conditions. The Mexican students' responses were the same to both types of items. The distinction was not a useful one in this instance.

Secondary findings within the study seem to warrant only brief comment. It does not seem surprising that the Mexican university students scored in a more internal direction than their labor worker compatriots. As young, upwardly mobile students, it follows that they would be more convinced of the efficacy of their actions than their less advantaged compatriots. That this difference would be marked in Mexico and not in the United States may be accounted for by the fact that being a university student in Mexico is a considerably more unique achievement than in this country. Two of us have argued on the basis of other data (Cole & Cole, 1977) that an internal locus of control will be particularly marked in an individual who has taken a counternormative step toward self-improvement. Pursuit of a

university degree in business administration is probably considerably more unusual in Mexico than in the United States.

The greater internality of the American students on the Politics subscale, when compared to scores on that scale for the American laborers, probably can be explained by the same rationale as is offered for the differences between the Mexican students and the Mexican laborers. (The American sample was a pre-Watergate sample.) On the other hand, both American student groups were significantly more externally oriented on the Respect subscale than the American laborers. This difference becomes more easily understood, however, if attention is directed to the content of the four items comprising this scale. Two of them deal directly with the question of how much assurance one can have that one is liked and accepted by one's peers. It is on these two items that the American students showed their strongest external orientation, and thus it seems likely that the external orientation on the Respect subscale obtained from these two American student groups reflects their status as young people still unsure of themselves in peer relationships, an insecurity much less evident in the Mexican students.

Results on the Leadership subscale were inconsistent, with one American group differing from the laborers and the other not. No explanation is offered.

It is of interest to note that the Irish and Mexican student groups share the common property of scoring most externally on the Politics subscale, something not found in the U.S. and West German samples. The political situations facing the Mexican and Irish students at the time of the testing may be of considerable relevance to this finding. The students at the Mexican university were at a school where there was a high degree of political unrest and opposition to the national government and where the data collection was held up several weeks while the school was closed by government order due to fear of student uprising. The Irish students were tested during a time of great tension between the political factions of Ireland, where acts of political terrorism were part of their daily experience. In light of these observations, it

is hardly surprising that these two groups showed more externality on the Politics subscale, and indeed one may wonder that the scores were not more external than they were.

Returning, however, to the chief thrust of this study, evidence to support a stereotype of a Mexican factory worker, a Mexican university student, or a Chicano high school senior as fatalistic, believing that his own actions are irrelevant to personal outcomes, is almost totally lacking. Instead, these groups appear equally or more internal in perceived locus of control than their American counterparts or other groups with whom they have been compared.

References

- Brislin, R. W. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1970, 1, 185-216.
- Cabrera, Y. A. A study of American and Mexican-American culture values and their significance in education. *Dissertation Abstracts*, 1964, 25, 309.
- Cole, D., & Cole, S. Counternormative behavior and locus of control. *Journal of Social Psychology*, 1977, 101, 21-28.
- Cross, W., & Maldonado, B. The counselor, the Mexican-American, and the stereotype. *Elementary School Guidance and Counseling*, 1971, 6, 27-31.
- Diaz-Guerrero, R. *Psychology of the Mexican: Culture and personality*. Austin: University of Texas Press, 1975.
- Garza, R. T., & Ames, R. E. A comparison of Anglo and Mexican-American college students on locus of control. *Journal of Consulting and Clinical Psychology*, 1974, 42, 919.
- Gilbert, G. M. Stereotype persistence and change among college students. *Journal of Abnormal and Social Psychology*, 1951, 46, 245-254.
- Gurin, P., Gurin, G., Lao, R., & Beattie, M. Internal-external control in the motivational dynamics of negro youth. *Journal of Social Issues*, 1969, 35(3), 29-54.
- Heller, C. *Mexican-American youth: Forgotten youth at the crossroads*. New York: Random House, 1966.
- Hsieh, T. T., Shybut, J., & Lotsof, E. F. Internal versus external control and ethnic group membership: A cross-cultural comparison. *Journal of Consulting and Clinical Psychology*, 1969, 33, 122-124.
- Justin, N. Culture, conflict, and the Mexican-American achievement. *School and Society*, 1970, 98, 27-28.
- Karlins, M., Coffman, T. L., & Walters, G. On the fading of social stereotypes: Studies in three gen-

- erations of college students. *Journal of Personality and Social Psychology*, 1969, 13, 1-16.
- Katz, D. E., & Braly, K. W. Racial stereotypes of 100 college students. *Journal of Abnormal and Social Psychology*, 1933, 28, 280-290.
- Levenson, H. Activism and powerful others: Distinctions within the concept of internal-external control. *Journal of Personality Assessment*, 1974, 38, 377-383.
- Lewis, O. *Five families: Mexican case studies in the culture of poverty*. New York: Basic Books, 1959.
- Mahler, I. A comparative study of locus of control. *Psychologia*, 1974, 17, 135-139.
- McGinnies, E., Norholm, L., Ward, C., & Bhanthumnavin, D. Sex and cultural differences in perceived locus of control among students in five countries. *Journal of Consulting and Clinical Psychology*, 1974, 42, 451-455.
- Morales, A. The collective preconscious and racism. *Social Casework*, 1971, 52, 285-293.
- Parsons, O., & Schneider, J. Locus of control in university students from eastern and western societies. *Journal of Consulting and Clinical Psychology*, 1974, 42, 456-461.
- Reitz, H., & Groff, G. Economic development and belief in locus of control among factory workers in four countries. *Journal of Cross-Cultural Psychology*, 1974, 5, 344-355.
- Rotter, J. Generalized expectancies for internal vs. external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).
- Schneider, J., & Parsons, O. Categories on the locus of control scale and cross-cultural comparisons in Denmark and the United States. *Journal of Cross-Cultural Psychology*, 1970, 1, 131-138.

Received October 19, 1977 ■

A Multiple-Component Treatment Approach to Smoking Reduction

Charles H. Elliott and Douglas R. Denney
University of Kansas

A package treatment program was designed to reduce cigarette smoking, and its effectiveness was compared with a single treatment condition (rapid smoking), a nonspecific treatment condition, and an untreated control condition. Following the treatment and posttesting sessions, another factor was introduced. One third of the subjects in each of the three treatment conditions were randomly assigned to specific booster (i.e., additional rapid-smoking) sessions, nonspecific booster sessions, or no booster sessions. Since the principal issue in the treatment of smoking is the maintenance rather than the induction of change, emphasis was placed on follow-up smoking levels 3 months and 6 months after the termination of treatment. The package condition was shown to produce substantially higher abstinence rates (45%) and lower percentages of baseline smoking (41%) after 6 months than the other treatment and control conditions. No reliable effects due to booster sessions were found.

Several reviews of the smoking reduction literature (Bernstein, 1969; Bernstein & McAlister, 1976; Hunt & Belpalec, 1974; Hunt & Matarazzo, 1973; Lichtenstein & Danaher, 1976; McFall & Hammen, 1971) have reached the following conclusions: (a) Virtually any treatment program is capable of reducing smoking levels to 30% or 40% of baseline; (b) a return to about 75% of baseline is commonly observed from 3 to 6 months after treatment; (c) seldom more than 13% of the subjects in any treatment program are completely abstinent after a 3- to 6-month follow-up period; and (d) of those subjects who are abstinent at the end of treatment, less than one third manage to maintain non-smoking 3-6 months later. High relapse rates have been observed for a wide variety of behavioral techniques, including systematic desensitization (Pyke, Agnew, & Kopperud,

1966), rapid smoking (Lando, 1975, 1976; Lichtenstein, Harris, Birchler, Wahl, & Schmahl, 1973; Schmahl, Lichtenstein, & Harris, 1972; Sutherland, Amit, Golden, & Rosenberger, 1975), covert sensitization (Sachs, Bean, & Morrow, 1970) aversive conditioning (Berecz, 1972; Whitman, 1972), stimulus control (Bernard & Efran, 1972), contingency contracting (Lawson & May, 1970), and behavioral rehearsal (Steffy, Meichenbaum, & Best, 1970).

The striking uniformity in these results generated by diverse techniques led McFall and Hammen (1971) to hypothesize that nonspecific factors were responsible for most of the reported changes in smoking rates. They designed a treatment procedure that incorporated only nonspecific factors such as motivated volunteering, self-monitoring, expectancy, demand, and mild encouragement. The procedure resulted in reduction rates and abstinence figures comparable to the figures cited above. Lichtenstein and Keutzer (1971) have argued that since smoking responds so readily to nonspecific factors, at least temporarily, minimal or nonspecific treatment groups constitute more adequate controls than do untreated groups in smoking reduc-

This research was funded by a University of Kansas general research grant awarded to the second author.

Charles H. Elliott is now at East Central University, Ada, Oklahoma.

Requests for reprints should be sent to Douglas R. Denney, Department of Psychology, University of Kansas, Lawrence, Kansas 66045.

tion outcome studies. Unfortunately, most of the above studies have failed to include non-specific treatment groups in their design.

Although initial studies focused on the effectiveness of single treatment procedures, attention now appears to be shifting to broad-based package therapies for treating maladaptive behaviors such as smoking (Lichtenstein & Danaher, 1976), alcoholism (Denney, 1976), and obesity (Mahoney, Note 1). Two basic arguments can be cited in favor of package approaches to treatment. First, the multi-determined nature of these maladaptive behaviors is becoming increasingly apparent. As investigators learn to appreciate the complexity of these behaviors and the variety of functions that cigarettes, alcohol, and food can serve for each individual, the necessity of package approaches that use a variety of techniques becomes clear. Second, from a strategic point of view, it would seem advisable to devise a complex treatment program that achieves the desired results in terms of bringing about persistent changes in these maladaptive behaviors and then subsequently to perform analytical studies to discover the effective components operating within the package. The current research on the reduction of smoking lacks clear demonstrations of long-term effectiveness. Only after such effects have been produced can the isolation of effective component procedures be undertaken.

Lichtenstein and Danaher (1976) surveyed the initial package approaches to smoking reduction and concluded that their effectiveness was no greater than single approaches. Five additional package approaches have since appeared in the literature and have generally indicated more encouraging results. Best (1975) evaluated a treatment package that was tailored to subjects' locus of control scores and included rapid smoking, aversive conditioning, attitude change techniques, and a review of environmental determinants of smoking. Best reported smoking levels at 30% of baseline and abstinence rates of 50% 6 months after treatment. Pederson, Scrimgeour, and Lefcoe's (1975) treatment package included hypnosis, relaxation training, self-monitoring, rehearsal of alternative behaviors, self-management training, and discussion. These investigators found that 50% of their

subjects were abstinent 6 months after treatment. Delahunt and Curran (1976) developed a package combining self-control and negative practice procedures. A 6-month abstinence rate of 56% for the nine subjects in the package condition was reported. Lando (1977) reported a 76% abstinence rate 6 months after administering a treatment package that included aversion, contractual management, booster sessions, group contact, and support. The one exception to these more encouraging results that adheres to package approaches is a study by Danaher (1977). His package approach, which combined rapid smoking with a variety of self-control procedures, resulted in abstinence rates of 21% and smoking levels of 52% of baseline after a 13-week follow-up. Furthermore, the package treatment in this final study was actually less effective than a procedure involving rapid smoking alone.

Danaher's (1977) pessimistic findings regarding package approaches are especially disturbing, since his study corrected for a number of methodological weaknesses in the preceding studies. All four of the preceding studies used single therapists who were aware of the hypotheses under investigation, thus augmenting the chances of experimenter bias. Three of these studies (Best, 1975; Lando, 1977; Pederson et al., 1975) failed to include nonspecific treatment groups, and, as Lichtenstein and Keutzer (1971) have argued, these groups represent the preferable form of control for smoking reduction studies. The data collection procedures used in three of the studies (Best, 1975; Delahunt & Curran, 1976; Pederson et al., 1975) relied exclusively on subjects' own counting of the number of cigarettes consumed or on their estimated rates of smoking, with no systematic attempts to check on the accuracy of these data. Finally, Lando (1977) required subjects to attend additional sessions of aversive treatment if they returned to smoking after treatment, a practice that clearly could lead to inflated results at follow-up.

The present study was designed to evaluate a broad-based treatment package for the reduction of smoking and to compare its effectiveness with a single treatment procedure (rapid smoking), which constituted a major

component within the treatment package. Thus, the results of this study are directly comparable to those reported by Danaher (1977). Like Danaher's investigation, the present study incorporated several methodological refinements including (a) multiple therapists; (b) a nonspecific treatment procedure, including information and group discussion along with the usual nonspecific factors accompanying any treatment; and (c) checks on the accuracy of subjects' cigarette consumption data. The selection of components for the current treatment package was based on research examining the effectiveness of each component as a single treatment procedure for reducing smoking. In particular, the treatment package was designed to enhance the maintenance and generalization of treatment effects by emphasizing the use of general, self-initiated coping strategies.

A second purpose of the present study was to evaluate the use of booster sessions for improving the maintenance of smoking reduction. Although several investigators have recommended the use of booster sessions and have incorporated such sessions in their treatment procedures (Chapman, Smith, & Layden, 1971; Hunt & Matarazzo, 1973; Lando, 1975, 1977; Lichtenstein et al., 1973; Schmahl et al., 1972), only one controlled investigation of the effectiveness of booster sessions has been reported. Kopel (1974) found no differences between subjects who received rapid-smoking booster sessions following treatment procedures involving rapid smoking and subjects who had received no such booster sessions. The design of the present study allowed us to compare subjects who had received specific booster sessions consisting of additional rapid-smoking trials, subjects who had received nonspecific booster sessions consisting of additional information and encouragement, and subjects who had received no booster sessions.

Method

Subjects

Newspaper advertisements were used to recruit smokers from the community of Lawrence, Kansas. To be accepted into the study, subjects had to be smoking at least 10 cigarettes per day and be willing to pay a \$30 deposit that would be returned to them

gradually over the course of the study, contingent only on their attendance at treatment and follow-up sessions. Of the 108 persons who attended an introductory session, 69 complied with these requirements and agreed to participate in the study.

Because of possible health hazards associated with rapid smoking (Hauser, 1974; Lichtenstein, 1974), subjects were required to complete a medical history questionnaire and to have their weight and blood pressure measured. Five subjects were dropped from the study because of possible health complications. One additional subject was dropped later in the study because of failure to faithfully use the cigarette collection method.

Pretest and posttest data were therefore available for 63 subjects (29 males, 34 females). This sample had the following characteristics: average age = 29.4 years; average education = 15.7 years; average length of time smoking = 12.4 years; average estimate of daily smoking level = 27.0 cigarettes; and average daily number of cigarettes smoked during a 7-day baseline period = 19.6 cigarettes. The number of subjects in the package treatment, rapid-smoking treatment, nonspecific treatment, and untreated control conditions were 20, 19, 18, and 6 respectively.¹ Three additional subjects were lost during the 6-month follow-up period, 1 each from the rapid-smoking treatment, nonspecific treatment, and control conditions.

Pretesting

A pretest session was conducted 1 week prior to treatment. During this session, all subjects completed two questionnaires. The smoking history questionnaire was designed to provide information descriptive of the subject sample and to obtain the name and phone number of a close friend to help check on the accuracy of the cigarette collection method. The semantic differential scale was designed to assess subjects' attitudes toward smoking. The concept *cigarette smoking* was rated on 10 7-point scales comprised of highly evaluative bipolar adjectives (e.g., healthy-unhealthy; attractive-ugly; soothing-irritating; fragrant-rank).

After completing these questionnaires, subjects were given a stack of cloth pouches, similar in design to tobacco pouches, in which to store their cigarette butts. Subjects were explicitly told not to count their cigarettes but simply to drop their butts into the pouches, using a new pouch each day.² The next 7 days prior to the first treatment session con-

¹ Since the subjects in the three treatment groups were later divided into subgroups receiving specific booster sessions, nonspecific booster sessions, and no booster sessions, the number of subjects in the untreated control group was similar to those of the other nine cells in the complete design.

² Pilot testing had shown that subjects would follow these instructions without counting the number of butts in the pouch.

stituted the baseline period during which the first count of cigarette consumption was completed.

Treatment

Following the pretest session, male and female subjects were assigned separately to four conditions: package treatment, rapid-smoking treatment, nonspecific treatment, and untreated control. Each of the three treatment conditions was administered to groups composed of 6-9 subjects. The subjects in these three treatment conditions attended three treatment sessions per week for 3 weeks. All subjects continued to collect their cigarette butts and to turn them in periodically during the 3 weeks of treatment.

Five advanced undergraduate psychology majors (three males, two females), served as therapists. One female therapist was responsible for the nonspecific procedure, which included giving brief informative lectures about smoking, handing out educational materials, collecting cigarette pouches, and giving mild encouragement. This nonspecific procedure was conducted for all subjects at the start of each treatment session, after which the subjects went to other rooms to receive their specific treatment procedures. Accordingly, throughout the study this therapist was unaware of the treatment conditions to which subjects had been assigned. The remaining four therapists were unaware of the hypotheses under investigation, and the treatment procedures they dispensed were highly standardized, with many components being presented by tape recorders.

At the first treatment session, subjects were told that they could attempt to quit smoking either immediately or gradually. The importance of accurate data collection was emphasized, and subjects were told that lie detector tests would be used at the end of the study to corroborate their conscientious use of the pouches. A brief description of each condition follows.³

Package treatment. In addition to the nonspecific procedure, eight component procedures were included in the package treatment. A rapid-smoking procedure (Lichtenstein et al., 1973) required subjects to smoke every 6 sec until they were unable to continue. Two rapid-smoking trials were presented during each treatment session. In the *applied relaxation* procedure (Chang-Liang & Denney, 1976), subjects practiced relaxation exercises while seated in a private cubicle listening to tape-recorded instructions. Subjects were also instructed in the application of relaxation as a coping skill during times when they felt a desire to smoke. The *covert sensitization* procedure (Gerson & Lanyon, 1972) encompassed both aversive scenes and relief scenes. In the former, subjects visualized themselves starting to smoke a cigarette in a particular setting, becoming nauseous, and vomiting. In the relief scene, subjects imagined themselves turning away from cigarettes and making some covert self-assertion ("I can control this habit"), whereupon the imagined feelings of nausea subsided. In the *systematic desensitization* procedure (Gerson & Lanyon, 1972), subjects were instructed to relax

and to imagine scenes involving situations that commonly elicited smoking. Six hierarchically ordered scenes were used in this procedure. The scenes had been selected on the basis of responses from a previous sample of smokers who responded to a survey concerning settings that commonly elicit smoking behavior. In the *self-reward and punishment* procedure (Thoreson & Mahoney, 1974), subjects completed contractual forms describing rewards and punishers that they would deliver to themselves contingent on their reaching or not reaching a self-selected smoking rate. In *cognitive restructuring* (Reed & Janis, 1974), subjects listed typical rationalizations as to why it was difficult for them to stop smoking. Later they were taught to formulate alternative ways of thinking about each rationalization and to use new self-verbalizations whenever tempted to smoke. *Behavioral rehearsal* (Chapman et al., 1971) involved practicing both verbal responses for turning down cigarettes in a variety of social settings and nonverbal behaviors (e.g., rubbing two coins together) as substitute activities during times when one was tempted to smoke. In *emotional role playing* (Janis & Mann, 1965), subjects prepared and then enacted scenes in which they learned that they had lung cancer and had to inform loved ones of their impending death.

In general, the package treatment introduced a wide variety of procedures rather than exhaustively dealing with each one. Table 1 illustrates the order of presentation and the approximate time devoted to each component procedure in the package condition during the nine treatment sessions.

Rapid smoking. This treatment was patterned after the rapid-smoking treatments described by Lichtenstein (Lichtenstein et al., 1973). Following the nonspecific procedure, subjects in this condition received two rapid-smoking trials conducted like those described for the package treatment. This sequence was repeated during each of the nine treatment sessions.

Nonspecific treatment. Subjects in this condition received the standard nonspecific procedure, including lectures, educational materials, mild encouragement, and data collection. To equate this condition with the preceding two conditions in terms of both time and plausibility, the subjects engaged in non-directive discussion for about 45 minutes following the standard nonspecific procedure.

Untreated control. Subjects in this group collected their cigarette butts and engaged in the pretest, posttest, and follow-up sessions, but they received no intervening treatment. They were told that they could use any of their own efforts to quit smoking during the data collection period.

Posttesting

The posttest session was conducted 1 week after the last treatment session. Subjects turned in the

³ More complete descriptions of each treatment condition are available from the second author on request.

Table 1
Package Treatment Group

Session	Treatment component	Approximate duration (min)
1	Nonspecific factors	10
	Treatment rationale and ratings	10
	Applied relaxation	20
	Rapid smoking	15
2	Nonspecific factors	10
	Applied relaxation	15
	Rapid smoking	15
	Self-reward training	10
3	Nonspecific factors	10
	Applied relaxation	15
	Rapid smoking	15
	Self-punishment training	10
4	Nonspecific factors	10
	Covert sensitization	20
	Instructions for emotional role playing	15
	Rapid smoking	15
5	Nonspecific factors	10
	Covert sensitization	20
	Emotional role playing	30
	Rapid smoking	15
6	Nonspecific factors	10
	Covert sensitization	15
	Emotional role playing	30
	Rapid smoking	15
7	Nonspecific factors	10
	Desensitization	20
	Lecture and distribution of cognitive restructuring questionnaire	10
	Rapid smoking	15
8	Nonspecific factors	10
	Desensitization	15
	Cognitive restructuring	25
	Rapid smoking	15
9	Behavior rehearsal	10
	Nonspecific factors	10
	Desensitization	15
	Behavior rehearsal	30
	Rapid smoking	15

cigarette butts collected over the 7 days subsequent to their last treatment and once again completed the semantic differential scale. Pouches for the week prior to the 3-month follow-up session were distributed.

Booster Sessions

Following the posttest session, subjects in each of the three treatment conditions were randomly divided into three booster conditions. Subjects as-

signed to the *specific booster* condition received three booster sessions, each of which included a brief refresher lecture, mild encouragement, and two additional rapid-smoking trials. Those in the *nonspecific booster* condition received three booster sessions consisting simply of the refresher lectures and mild encouragement. Those in the *no-booster* condition received no contact between the posttest and the first follow-up sessions. Booster sessions were conducted during the 1st, 3rd, and 5th weeks after the posttest session.

3-Month and 6-Month Follow-Up

Three months after the posttest session, subjects were contacted by phone and were reminded to begin using their collection pouches for the next 7 days. They were also asked to estimate their daily smoking level. One week later the subjects returned to the laboratory to turn in their pouches and complete the semantic differential scale.

The second follow-up, conducted 6 months after the posttest, was identical to the first. After the pouches had been collected and the questionnaire was completed, all subjects were debriefed and their deposits were returned. Untreated subjects and those who were not satisfied with their progress in the treatment conditions were offered additional individual treatment.

Data Accuracy Checks

It was possible to obtain the help of a friend for 56 of the subjects. The friend agreed to observe whether the pouches were being used conscientiously by the subject. These friends were contacted once during the treatment sessions and once during each follow-up session. Their reports resulted in the exclusion of only 1 subject's data.

As an additional check on the data, 55 subjects were contacted by phone 4 days after the second follow-up session. Using a nonreactive measure inspired by Sushinsky (1972), a confederate attempted to conduct a marketing survey in which the subject was asked to report on the quantity of several nonfood grocery items consumed. One item in the survey was cigarettes. Forty-six subjects agreed to take part in the survey. Two were later eliminated because they suspected a connection to the smoking study. In no instance did the estimates given by the remaining subjects vary more than 20% from either the smoking level estimates or the actual pouch count collected during the second follow-up session. Most importantly, not one of the subjects who claimed to be abstinent admitted to smoking during the nonreactive call.

Results

Preliminary Analyses

A number of preliminary analyses of variance were performed on the data, producing

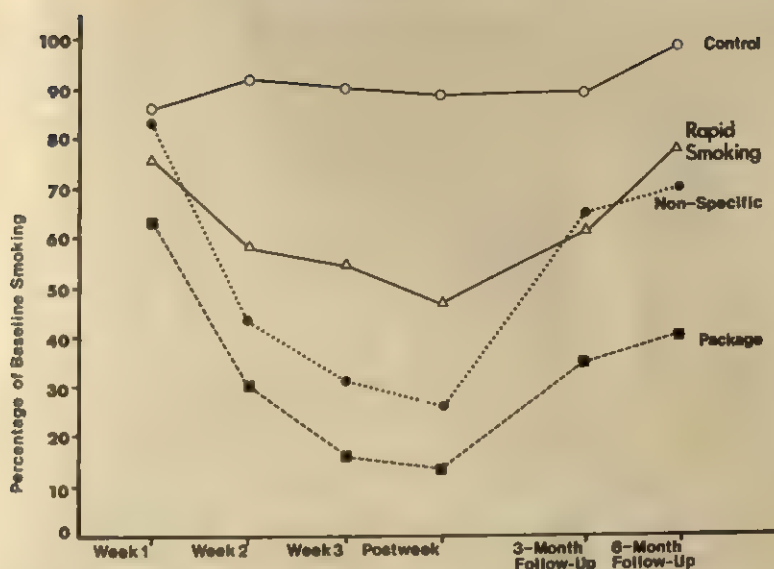


Figure 1. Cigarette consumption as a percentage of the baseline level of smoking.

the following results: (a) No significant differences were found among subjects in the four conditions (i.e., package, rapid smoking, nonspecific, untreated control) in terms of age, years of education, years of smoking, estimated motivation to quit, estimated daily smoking level, or locus of control; (b) no significant initial differences existed among subjects in the four conditions in terms of any of the repeated measures, that is, baseline smoking level or attitudes toward smoking; (c) no significant differences existed among subjects in the three treatment condition (i.e., package, rapid smoking nonspecific) in terms of expectancy of success as measured by a five-item expectancy questionnaire administered during the first treatment session; (d) no significant main effects for sex and no significant Sex \times Condition interactions existed on any of the measures described above; and (e) no significant main effects for sex and no significant Sex \times Condition interactions were found in terms of outcome on the two repeated measures. Accordingly, the four conditions were considered equal at the start of treatment, and sex was eliminated as a factor in all subsequent analyses.

Cigarette Consumption

Percentage of each subject's baseline smoking was used to reflect changes in cigarette consumption. These percentages were considered a more appropriate gauge of treatment outcome than actual numbers of cigarettes smoked and could easily be compared with other treatment studies (McFall & Hammen, 1971).⁴ Figure 1 depicts the changes that occurred in percentage of baseline smoking over the course of treatment, posttest, and follow-up sessions.

A 4 (treatment) \times 4 (trials) analysis of covariance was performed on the percentage of baseline smoking scores for the 3 treatment weeks and the posttest week. The average daily number of cigarettes smoked during the baseline week served as the covariate. A significant Treatment \times Trials interaction was found, $F(9, 177) = 5.96, p < .001$. A series of one-way analyses of covariance and Duncan pairwise comparison tests were performed to further analyze this interaction. Each of the treatment conditions resulted in significant decreases in smoking across trials

⁴ Number of cigarettes smoked was also analyzed and yielded results that were equivalent to those for the percentage of baseline smoking.

(all $ps < .001$), whereas the untreated control condition resulted in no change ($F < 1$). At the time of the posttest, the subjects in the three treatment conditions were smoking significantly less than those in the untreated control condition (all $ps < .005$). Furthermore, the package condition resulted in significantly less smoking than the rapid-smoking condition. The difference between the package condition and the nonspecific condition was not significant, but it did favor the package condition.

A 3 (treatment) \times 3 (booster) \times 3 (trials) analysis of covariance was performed on the percentage of baseline smoking scores for the three treatment conditions during the posttest, 1st follow-up, and 2nd follow-up weeks. Again, the average daily number of cigarettes smoked during baseline served as the covariate. This analysis revealed no significant main effect or interactions involving booster as a factor. Accordingly, the booster factor was dropped, and a 4 (treatment) \times 3 (trials) analysis of covariance was performed on posttest, first follow-up, and second follow-up data, with the untreated control condition being added to the design. Significant main effects for treatments, $F(3, 55) = 6.59$, $p < .001$, and for trials, $F(3, 55) = 14.85$, $p < .001$, were found. Specific comparisons revealed that subjects in each of the three treatment conditions showed a significant return in the direction of baseline from the posttest to the subsequent follow-up weeks (all $ps < .005$). Subjects in the untreated control condition showed no change during this period ($F < 1$). At the time of the first follow-up session, subjects in the package condition were smoking less than subjects in the rapid-smoking ($p < .07$), nonspecific ($p < .05$), and untreated control conditions ($p < .005$). At the time of the second follow-up session, subjects in the package condition were still smoking less than those in the rapid-smoking ($p < .07$) or untreated control conditions ($p < .005$). Although the difference between the package condition and the nonspecific condition was 28.6%, this difference failed to attain significance. Nonetheless, at the time of both the first and the second follow-up sessions, the package condi-

tion was the only treatment condition that differed significantly from the untreated control condition.

In addition to the percentage of baseline smoking data, the proportions of completely abstinent subjects in each condition were examined. Chi-square analyses demonstrated significant differences in the proportion of abstainers in the various conditions at the posttest session, $\chi^2(3) = 11.1$, $p < .02$, and at the second follow-up session, $\chi^2(3) = 8.5$, $p < .04$. At posttest, the package condition contained a substantially larger proportion of abstainers (65%) than the control condition (0%, $p < .005$), the rapid-smoking (26%, $p < .02$), or the nonspecific condition (33%, $p < .06$). Similarly, at the time of the second follow-up, the package condition contained a larger proportion of abstainers (45%) than the control condition (0%, $p < .06$), the rapid-smoking condition (17%, $p < .06$), or the nonspecific condition (12%, $p < .03$). It is clear that most of the differences between conditions that occurred in the percentage of baseline smoking data can be accounted for in terms of the proportion of abstainers found in each of the conditions.

Attitudes Toward Smoking

A 3 (treatment) \times 3 (booster) \times 3 (trials) analysis of covariance was performed on the posttest, first follow-up, and second follow-up scores of the semantic differential scale, with the corresponding pretest scores serving as the covariate. This analysis failed to reveal any significant main effects or interactions involving booster sessions. Accordingly, the booster factor was dropped, the untreated control condition was added to the treatment factor, and a 4 (treatment) \times 3 (trials) analysis of covariance was performed on the scores from the semantic differential scale. This analysis revealed a significant main effect for treatment, $F(3, 55) = 12.06$, $p < .001$. Duncan pairwise comparisons revealed that the subjects in the package condition held significantly more negative attitudes toward cigarette smoking than subjects in all other conditions at the time of the posttest (all $ps < .001$), with no differences occurring

among the other three conditions. At the first follow-up session, subjects in the package condition continued to express more negative attitudes than subjects in the control condition ($p < .001$), the nonspecific condition ($p < .005$), or the rapid-smoking condition ($p < .07$). At the time of the second follow-up session, subjects in the package condition were still evaluating cigarette smoking somewhat more negatively than those in the control condition ($p < .005$), the nonspecific condition ($p < .08$), and the rapid-smoking condition ($p < .11$). All other comparisons failed to approach significance.

Discussion

Six months after treatment, the package treatment had produced a percentage of baseline smoking of 41% and an abstinence rate of 45%. These results are substantially better than figures reported for single treatment studies (McFall & Hammen, 1971) and also better than figures reported by Danaher (1977) in his package treatment study. These results are somewhat lower than those of other recent investigations using combinations of techniques (Best, 1975; Delahunt & Curran, 1976; Lando, 1977; Pederson et al., 1975). However, as indicated earlier, these latter studies suffered from several methodological weaknesses that may have inflated their results.

The present study contained a number of safeguards to improve on the validity of its findings. Experimenter bias was minimized by using tape-recorded treatment components and multiple therapists who were blind to the hypotheses of the study and the conditions being compared. In addition, a number of checks were used to help assure the accuracy of subjects' cigarette consumption data. Subjects were informed that they would receive lie detector tests, informants were used to confirm subjects' smoking status, and a disguised market survey was used at the end of the study.

The package condition also demonstrated highly favorable results when compared with the other conditions in the present study. After 6 months, the package treatment was the only condition with a percentage of base-

line smoking significantly lower than the untreated control condition. In addition, this percentage figure was 38% below that of the rapid-smoking condition and 29% below that of the nonspecific condition. Finally, the abstinence rate for the package condition at 6 months was substantially greater than the abstinence rate for each of the other conditions (all $ps < .06$).

Clearly, the package treatment demonstrated effectiveness beyond what could be attributed to nonspecific factors such as motivated volunteering, structure, self-monitoring, information, and encouragement. Borckovec (1973) has argued that placebo groups or nonspecific treatment groups seldom constitute adequate controls for expectancy, since the subjects in these groups rarely hold as favorable a set of expectancies as do those in the valid treatment groups. In the present study, however, an expectancy questionnaire was administered during the first treatment sessions, after subjects had heard descriptions of the treatments that they were about to receive. Subjects in all three treatment conditions held comparable initial expectancies, suggesting that differential expectancy at the start of treatment does not account for the differences between the package and the nonspecific conditions.

The present design did not permit the identification of the effects attributable to each of the component procedures within the package treatment. However, reports on a follow-up questionnaire by a number of the subjects indicated that cognitive restructuring and emotional role playing were particularly useful in aiding them in making the "real" decision to quit and that relaxation training and behavioral rehearsal were useful as coping strategies once that decision was reached. In contrast, the rapid-smoking component received almost no endorsement by subjects. In addition, the highly aversive nature of this procedure was clearly demonstrated. Six subjects actually vomited during rapid-smoking trials. Most other subjects reported extreme physical discomfort, including dizziness, headaches, sweating, and nausea, with effects lasting well over an hour after rapid-smoking trials.

In spite of its aversiveness, rapid smoking did not prove to be a very effective procedure. Subjects in the rapid-smoking condition achieved a percentage of baseline smoking of 73% and an abstinence rate of 17% after 6 months, figures that are almost identical to the results attained by McFall and Hammen (1971) using nonspecific treatment. Clearly, the results of our study stand in sharp contrast to those of Danaher (1977), both in terms of the effectiveness of our package treatment and the lack of effectiveness of rapid smoking. Further research is needed to explain these differences, but for the present, two points are noteworthy. First, the results obtained for our treatment package are much more comparable to those of other treatment packages (Best, 1975; Delahunt & Curran, 1976; Pederson et al., 1975) than is the case for Danaher's study. Second, other recent investigations (Lando, 1975, 1976; Sutherland et al., 1975) have begun to cast doubt on the effectiveness of rapid smoking.

The package treatment might be improved by strengthening the cognitive restructuring, emotional role-playing, relaxation, and behavioral rehearsal components and eliminating the rapid-smoking trials. It would also seem advisable to include more procedures aimed explicitly at teaching controlled smoking. Almost all of the success in the package condition was due to relatively large proportion of abstinent subjects in that condition. Little was accomplished by way of teaching nonabstinent subjects to smoke at reduced, medically safer levels. Controlled-smoking procedures addressed explicitly to those who are unwilling to quit might represent an improvement over the current package treatment. The issues regarding abstinence and controlled smoking are comparable to those currently being debated in the area of alcoholic treatment (e.g., Vogler, Compton, & Weissbach, 1975).

Finally, the results of the present study replicate those of Kopel (1974) in demonstrating the lack of effectiveness of booster sessions. In spite of several investigators' (e.g., Hunt & Matarazzo, 1973; Lando, 1975) remarks that booster sessions might

insure continued "immunization" among non-smokers while rescuing backsliders, the present study revealed no effects for booster sessions in terms of cigarette consumption or self-report measures of attitudes toward smoking. In considering these results, it should be noted that the booster sessions in the current study did not represent complete replications of the original treatments that may have attenuated their potential effectiveness.

Reference Note

1. Mahoney, M. J. *Clinical issues in self-control training*. Paper presented at the 81st annual meeting of the American Psychological Association, Montreal, August 1973.

References

- Berecz, J. Modification of smoking behavior through self-administered punishment of imagined behavior: A new approach to aversion therapy. *Journal of Consulting and Clinical Psychology*, 1972, 38, 244-250.
- Bernard, H. S., & Efran, J. S. Eliminating versus reducing smoking using pocket timers. *Behaviour Research and Therapy*, 1972, 10, 399-401.
- Bernstein, D. A. Modification of smoking behavior: An evaluative review. *Psychological Bulletin*, 1969, 71, 418-440.
- Bernstein, D. A., & McAlister, A. The modification of smoking behavior: Progress and problems. *Addictive Behaviors*, 1976, 1, 89-102.
- Best, J. A. Tailoring smoking withdrawal procedures to personality and motivational differences. *Journal of Consulting and Clinical Psychology*, 1975, 43, 1-8.
- Borkovec, T. D. The role of expectancy and physiological feedback in fear research: A review with special reference to subject characteristics. *Behavior Therapy*, 1973, 4, 491-505.
- Chang-Liang, R., & Denney, D. R. Applied relaxation as training in self-control. *Journal of Counseling Psychology*, 1976, 23, 183-189.
- Chapman, R. F., Smith, J. W., & Layden, L. A. Elimination of cigarette smoking by punishment and self-management training. *Behaviour Research and Therapy*, 1971, 9, 255-264.
- Danaher, B. G. Rapid smoking and self-control in the modification of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1068-1075.
- Delahunt, J., & Curran, J. P. Effectiveness of negative practice and self-control techniques in the reduction of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1976, 44, 1002-1007.

- Denney, D. R. Behavioral approaches to the treatment of alcoholism. In G. Goldstein & C. Neuringer (Eds.), *Experimental studies in alcoholism*. Cambridge, Mass.: Ballinger, 1976.
- Gerson, P., & Lanyon, R. I. Modification of smoking behavior with an aversion desensitization procedure. *Journal of Consulting and Clinical Psychology*, 1972, 38, 399-402.
- Hauser, R. Rapid smoking as a technique of behavior modification: Caution in selection of subjects. *Journal of Consulting and Clinical Psychology*, 1974, 42, 625.
- Hunt, W. A., & Bespalec, D. A. An evaluation of current methods of modifying smoking behavior. *Journal of Clinical Psychology*, 1974, 30, 431-438.
- Hunt, W. A., & Matarazzo, J. D. Three years later: Recent developments in the experimental modification of smoking behavior. *Journal of Abnormal Psychology*, 1973, 81, 107-114.
- Janis, I. L., & Mann, L. Effectiveness of emotional role-playing in modifying smoking habits and attitudes. *Journal of Experimental Research in Personality*, 1965, 1, 84-90.
- Kopel, S. A. *The effects of self-control, booster sessions, and cognitive factors on the maintenance of smoking reduction*. Unpublished doctoral dissertation, University of Oregon, 1974.
- Lando, H. A. A comparison of excessive and rapid smoking in the modification of chronic smoking behavior. *Journal of Consulting and Clinical Psychology*, 1975, 43, 350-355.
- Lando, H. A. Aversive conditioning and contingency management in the treatment of smoking. *Journal of Consulting and Clinical Psychology*, 1976, 44, 312.
- Lando, H. A. Successful treatment of smokers with a broad-spectrum behavioral approach. *Journal of Consulting and Clinical Psychology*, 1977, 45, 361-366.
- Lawson, D. M., & May, R. B. Three procedures for the extinction of smoking behavior. *Psychological Record*, 1970, 20, 151-157.
- Lichtenstein, E. Rapid smoking as a technique of behavior modification: Lichtenstein replies. *Journal of Consulting and Clinical Psychology*, 1974, 42, 625.
- Lichtenstein, E., & Danaher, B. G. Modification of smoking behavior: A critical analysis of theory, research, and practice. In M. Herson, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 3). New York: Academic Press, 1976.
- Lichtenstein, E., & Keutzer, C. S. Modification of smoking behavior: A later look. In R. D. Rubin, H. Fensterheim, A. A. Lazarus, & C. M. Franks (Eds.), *Advances in behavior therapy*. New York: Academic Press, 1971.
- Lichtenstein, E., Harris, D. E., Birchler, G. R., Wahl, J. M., & Schmah, D. P. A comparison of rapid smoking, warm, smoky air, and attention placebo in the modification of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1973, 40, 92-98.
- McFall, R. M., & Hammen, C. L. Motivation, structure, and self-monitoring: Role of non-specific factors in smoking reduction. *Journal of Consulting and Clinical Psychology*, 1971, 37, 80-86.
- Pederson, L. L., Scrimgeour, W. G., & Lefcoe, N. M. Comparison of hypnosis plus counseling, counseling alone, and hypnosis alone in a community service smoking withdrawal program. *Journal of Consulting and Clinical Psychology*, 1975, 43, 920.
- Pyke, S., Agnew, N. M., & Kopperud, J. Modification of an overlearned maladaptive response through a relearning program: A pilot study on smoking. *Behaviour Research and Therapy*, 1966, 4, 197-203.
- Reed, H. D., Jr., & Janis, I. L. Effects of a new type of psychological treatment on smokers' resistance to warnings about health hazards. *Journal of Consulting and Clinical Psychology*, 1974, 42, 748.
- Sachs, L. B., Bean, H., & Morrow, J. E. Comparison of smoking treatments. *Behavior Therapy*, 1970, 1, 465-472.
- Schmah, D. P., Lichtenstein, E., & Harris, D. E. Successful treatment of habitual smokers with warm, smoky air and rapid smoking. *Journal of Consulting and Clinical Psychology*, 1972, 38, 105-111.
- Steffy, R. A. Meichenbaum, D., & Best, J. S. Aversion and cognitive factors in the modification of smoking behavior. *Behaviour Research and Therapy*, 1970, 8, 115-125.
- Sushinsky, L. W. Expectation of future treatment, stimulus satiation, and smoking. *Journal of Consulting and Clinical Psychology*, 1972, 39, 343.
- Sutherland, A., Amit, Z., Golden, M., & Rosenberger, Z. Comparison of three behavioral techniques in the modification of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1975, 43, 443-447.
- Thoreson, C. E., & Mahoney, M. J. *Behavioral self-control*. New York: Holt, Rinehart & Winston, 1974.
- Vogler, R. E., Compton, J. V., & Weissbach, T. A. Integrated behavior change techniques for alcoholics. *Journal of Consulting and Clinical Psychology*, 1975, 43, 233-243.
- Whitman, T. L. Aversive control of smoking behavior in a group setting. *Behaviour Research and Therapy*, 1972, 10, 97-104.

Geriatric Patients Who Improve in Token Economy and General Milieu Treatment Programs: A Multivariate Analysis

Brian L. Mishara

University of Massachusetts at Boston

Chronic geriatric mental hospital patients were randomly assigned to token economy and general milieu programs. In each program, staff received the same amount of training, and the physical environments were identical. Results showed significant decreases over 6 months on both wards in frequency of bizarre and unusual behaviors. Incontinence decreased immediately after transfer to the treatment wards, and there were changes over 6 months in the amount of staff care given to patients. Multiple discriminant analyses indicated that in each ward different constellations of pretreatment characteristics discriminated between patients who improved and those who did not improve in the frequency of bizarre and unusual behaviors. In the token economy, improved patients can be characterized as less "institutionalized," in better physical condition, and actively exhibiting their troubles. In the general milieu, improved patients can be discriminated by less responsiveness to an interviewer. Results are interpreted in view of the different characteristics of each program.

Most research on the rehabilitation of chronic elderly mental hospital patients has focused on finding and evaluating one potentially useful treatment method rather than comparing techniques or trying to determine which elders benefit from which intervention programs. Two promising intervention techniques, token economy and general milieu programs, have been shown to be effective, but not all clients improve under treatment. This poses a dilemma for the practitioner who must choose an appropriate intervention method for particular clients. Often the

choice is made on the basis of ideological preferences rather than empirical data regarding which client characteristics relate to improvement under each condition. This article offers preliminary indications based on multiple discriminant analyses of which pretreatment client characteristics discriminate between those who improved and those who did not improve in a token economy and a general milieu treatment program.

There have been few attempts to discriminate between geriatric clients who improve under treatment and those who do not. Kleban, Lawton, Brody, and Moss (1976) used multivariate techniques to determine which characteristics of their subjects best predicted improvement in either an individually planned treatment program or a standardized treatment procedure. Their results indicated that age was the single most consistent predictor of success in both groups, with the younger subjects more likely to improve. Those who improved most in the individualized program tended to be more self-sufficient and active, with the social skills necessary to profit from their individualized contact with staff.

This study was conducted at Northville State Hospital, Northville, Michigan. It is based in part on a paper presented at the annual meeting of the American Psychological Association, San Francisco, August 1977.

The author wishes to thank Barbara Robertson, John Moir, and the other Northville staff whose involvement in the implementation of these programs was essential. Also, thanks are due to Derek Hybels and John Murphy for help in data analysis and A. Harvey Baker, Larry Gaines, and Robert Kastenbaum for their comments on earlier drafts of this article.

Requests for reprints should be sent to Brian L. Mishara, Department of Psychology, University of Massachusetts, Boston, Massachusetts 02125.

Sherwood, Morris, and Barnhart (1975) used the results of a discriminant function analysis to develop a system for assigning elderly people to appropriate residential settings. Their system was designed to differentiate elders who need the greater care offered by institutionalization from those who require less supervised settings or who can live independently or semi-independently. After further study and cross-validations, their approach may prove useful in deciding who should be institutionalized and who would be better off living independently.

The Sherwood et al. (1975) and Kleban et al. (1976) studies assessed characteristics of people who would benefit from more or less individual care. The present study is based on treatment conditions in which the amount of individual care was designed to be equivalent under both conditions. The major difference between the programs was their derivations from quite different theoretical perspectives. The token economy program, in the behaviorist tradition, involved careful application of operant reinforcement principles. On the other hand, the general milieu program provided the same reinforcements freely in the tradition of humanistic psychology and current research in environmental psychology.

Milieu treatment methods have recently gained attention for their potential to rehabilitate long-term institutionalized geriatric patients (e.g., Bok, 1971; Gottesman, 1973; Gottesman, Quarterman, & Cohn, 1973). Their findings have been generally accepted—An improved milieu is more conducive to positive therapeutic change than impoverished custodial care settings. Although these findings seem almost self-evident, findings from research on milieu treatment provided the impetus for practitioners to rehabilitate many chronic geriatric mental patients who were previously denied treatment, because it was thought that treatment attempts would be a waste of time.

Token economy programs that utilize operant conditioning principles have been quite successful with chronic mental patients (e.g., Ayllon & Azrin, 1968; Ayllon & Michel, 1959; Krasner, 1971). A number of

the token economy studies included elderly subjects (Atthowe & Krasner, 1968; Lee, 1969; Mueller & Atlas, 1972). Although token economies for just the elderly have been rare, behavior modification techniques have been increasingly accepted as a useful method for geriatric intervention (see Hoyer, Mishara, & Riedel, 1975).

In general, available data suggest that token economy programs are better than control situations in which no treatment is given (Gripp & Magaro, 1971; Maley, Feldman, & Ruskin, 1973; Shean & Zeidberg, 1971). Greenberg, Scott, Pisa, and Friesen (1975) found that a milieu treatment program combined with token economy was more effective than a token economy alone. However, the review by Gripp and Margaro (1974) suggested that it is still unclear which variables affected the outcome of token economy programs.

In this study discriminant function analysis was chosen, since the discrimination made parallels the real-life decisions faced in rehabilitation settings. With the development of different kinds of treatment for geriatric patients, hospital staff are faced with the problem of choosing between the various rehabilitation approaches that are currently available. The discriminant function approach may help to identify variables that may affect the outcome of such choices. Also, because of their diagnoses as suffering from incurable degenerative brain disease, it is frequently expected that no such patients would benefit from rehabilitation efforts. Their medical prognosis was continued decline. Given this expectation, *any* improvement sustained over 6 months could be an important indication of rehabilitation potential, hence the choice of two discrete populations, the "improved" and "unimproved." It was decided a priori that since these were mental hospital patients, the main indication of improvement would be decreases in the frequency of bizarre and unusual behaviors. Secondary measures of improvement were reduced frequencies of incontinence, increased interpersonal communication, and reduced staff care given for personal hygiene and self-care.

Table 1
Examples of Ways of Earning Tokens

Behavior	No. tokens
Social	
Helping another resident	6
Talking to another resident	4 ^a
Starting an activity by yourself	4
Starting an activity with others	8 ^b
Self-care	
Bathe self (including washing hair)	3
Dress self	— ^c
Comb hair and brush teeth	— ^c
Shave self or shine shoes	2
Especially neatly dressed in morning	1
Help do own hair	2
Ward work (partial list)	
Make bed	1
Mop	
Whole hall	7
One day room	3
Bathroom	2
After incontinence	3
Carry out trash	1
Dust one day room	1
Help distribute trays	1
Iron dresses	— ^d
Work off ward or make handicrafts	5 ^e
Wash ashtrays or medication cups	2
Stack chairs after meals	4

^a Per valid conversation.

^b For each person who participates.

^c 2 = without help; 1 = with a little help.

^d 2 = good job; 1 = sloppy job.

^e Per hour.

Method

Subjects

Eighty elderly people (40 men and 40 women) were selected randomly from the "medical" unit of a large state mental hospital to populate two new coed treatment wards, each with 20 males and 20 females. These people were on the medical unit because of their diagnoses of suffering from incurable chronic brain syndromes presumably associated with old age. Approximately half of the subjects were selected from those people who gave evidence of ability to communicate with others verbally. At the time of the original selection, only 31 subjects were communicative. Half of the communicative subjects were randomly assigned to each ward, and the remaining subjects were randomly divided between the wards.

The mean age was 68.8 years ($SD = 5.1$ years). The mean length of current hospitalization was 21.4 years ($SD = 14.7$ years). Most of the participants were previously involved in active rehabilitation programs on other units. These patients were the "failures." After failing to meet criteria of rehabilitation set up by these programs, people were transferred to the medical unit.

On the medical unit, only custodial care was available. A medical doctor made daily rounds, but there was no psychosocial treatment available—there were no psychologists, psychiatrists, or social workers involved actively in their treatment. The patients mostly wore hospital gowns in place of regular clothing, few ever left the wards, and the only available recreation was television.

Procedure

Identical numbers of nursing staff with equivalent training were assigned to each program. Nursing staff on all three shifts on each ward first participated in 13 weeks of training meetings for 3–4 hours a week. Training included discussions of how it feels to work with chronic elderly patients, demonstrations of how it feels to be disabled, and what it is like to work in a custodial hospital ward. The staff was instructed that the programs would involve (a) more activities and social stimulation, including better foods, radio-phonographs, social events, environmental decorations, better clothes, and so forth; (b) increased opportunity for patients to choose how to run their daily lives; and (c) more awareness of staff for whose benefit they were doing things: to satisfy their own needs or the needs of the patients.

During the last 7 weeks of the training period, staff on the general milieu ward participated in open-ended discussion, whereas staff on the token ward learned the operant conditioning procedures and theory involved in a token economy program.

Basically, the token economy program was modeled after that designed by Ayllon and Azrin (1968). Individuals were rewarded for desirable behaviors including social interaction with other patients, ward work, personal hygiene, and self-care activities. Table 1 lists some of the ways tokens could be earned. A detailed but simple bookkeeping system was used to keep track of all tokens given out and taken in, and the midnight shift audited the daily tallies and reported any discrepancies. Tokens could be exchanged for secondary reinforcements such as cigarettes, wine, permission to leave the ward, extra food, and other supplements. The staff met regularly to revise the token system and determine which behaviors should be rewarded.

The same secondary reinforcements that could be purchased by the token economy ward were available free to anyone on the general milieu ward who wanted them. Staff were instructed to increase environmental opportunities without any systematic reinforcement of behaviors. For example, in the

general milieu ward, food supplements such as coffee, wine, ice cream, and sandwiches were available free on a daily basis to anyone who wanted them. Whereas on the token economy ward these same foods were sold for tokens. On the general milieu ward, activities such as making handicrafts were available to anyone who wanted to participate, but on the token economy ward patients were paid by the hour for their work on these projects.

The physical environments on both wards were identical. The wards had duplicate floor plans and duplicate furniture at the start of the programs. On both wards efforts were made to acquire decorations and some more comfortable furniture, though the environments could still be regarded as quite "institutional" in character.

Prediction Measures

Before being transferred to the new research wards, data on all participants were gathered on the following measures, which were later entered into the discriminant analyses as the independent variables:

Psychotic Inpatient Profile. All participants were observed for 3 days by nursing staff using Lorr's Psychotic Inpatient Profile (PIP; Lorr & Vestre, 1969).

VIRO (Vigor Intactness Relationship Orientation). Research assistants rated subjects' interpersonal behavior using the VIRO scale developed for use with geriatric patients by Kastenbaum and Sherwood (1972). The VIRO scale consists of three scores: Presentation (based on the initial interpersonal behavior toward the interviewer), Interaction (based on interpersonal behavior during $\frac{1}{2}$ hour, and Orientation (which assesses orientation as to time, place, and person).

Staff evaluations of patient abilities. Staff who were familiar with the patients filled out questionnaires concerning observed dressing abilities and habits, personal hygiene level, physical condition (difficulties and illnesses), and behavioral disabilities.

Prediction Measure Reliabilities

As a check of the reliability of the PIP observations, on 10 occasions a research assistant spent 3 days observing on the ward in the course of other activities and filled out profiles on participants who were being observed by staff. There were few disagreements between ratings by staff and observers, and what little there was reflected a seemingly random pattern of errors of omission on the part of staff. To test the reliability of the VIRO, 20 elderly patients were interviewed twice within 2 weeks by different research assistants. The correlations between the two interviewers on the sections of VIRO were .94, .98, and .99.

Treatment Outcome Measures

The treatment outcome measures were based on 1-day observations by nursing staff, who recorded

the frequencies of occurrence of specific target behaviors. Observations were conducted in a random sequence, so that for 1 day every 2 weeks each patient was observed by a randomly assigned staff member on both the day and afternoon shifts. Staff conducted the observations during the course of their usual daily activities. Observations were from the month before the patients were transferred to research wards until 6 mo. after the programs began.

The main outcome measure was the frequency of bizarre and unusual behaviors. The behaviors included in this variable were defined on an individual basis after extensive observations. Typical types of behaviors included confused verbalizations (e.g., "My mother is coming today to take me to the zoo."), conversations or reports of conversations with people not actually present (e.g., "Jesus Christ just told me to throw peas on the ceiling."), self-injurious behaviors (e.g., scratching one's arm until it bleeds), unusual movements (e.g., "sweeping" the floor without a broom), and so forth. Secondary outcome measures were the frequency of occurrence of incontinence, the frequency of conversations with other patients, and the number of times nursing staff gave care or help in personal hygiene and patient self-care activities.

Treatment Outcome Measure Reliabilities

As a check of the reliability of these staff observations, outside experimenters occasionally spent 1 day on the ward observing the same behaviors in specific patients. Disagreements between staff and observer ratings were few, and they seemed to follow a random pattern. As a second check on the rating reliabilities, identical independent observations were conducted on the same day for the same patients by members of the day and afternoon shifts and were compared. As was expected, the day shift staff, who spent more waking-time contact with patients in general, reported more behaviors. The Spearman rank-order correlations (ρ) between observations on the two shifts ranged from .71 to .92 ($n = 37$ in the general milieu program, and $n = 36$ in the token economy), with the median correlation being .77. (All correlations were significant at $p < .01$.)

Statistical Analyses

General improvement according to staff observations was analyzed by sign tests on changes from before treatment to the end of 6 months on the rehabilitation wards. (See Siegel, 1956, pp. 68-75.) Sign tests were chosen for the primary analysis, since the data were not interval data.

For the multiple discriminant analyses, patients were classified as either having improved or as not having improved from before the start of the treatment program until 6 months after being transferred to the new treatment units. For each separate program a stepwise linear discriminant analysis was performed using the nine PIP scales, the three VIRO scales, the four staff evaluations of areas of patient ability, and age, sex, and total years of hospitalization

to determine which of these factors discriminated between the population of individuals who improved and those who did not improve. Supplementary multiple regression analyses were performed in order to gain an indication of the amount of variance accounted for by the measures.

Results

The first issue is whether or not either or both of the treatment programs had an effect on the primary target variable or any of the three secondary outcome measures. Given that these patients had a prognosis of further deterioration, it is not surprising that the majority showed no change from before the start of the programs to 6 months after the programs began. Most patients who changed decreased the frequency of bizarre and unusual behaviors (see Table 2). These improvements occurred in both the token economy (sign test, $p = .059$) and the general milieu (sign test, $p = .003$) programs, showing treatment effects.

Regarding the secondary outcome measures (see Table 2), there were significant improve-

ments on frequency of incontinence in the general milieu program (sign test, $p < .001$) and improvements in the token economy programs (sign test, $p = .062$). Care given by nursing staff was significantly reduced after 6 months only on the token economy ward (sign test, $p < .001$), and there was no significant improvement on frequencies of conversations with other patients on either ward.

Inspection of the data indicated that the declines in frequency of incontinence occurred entirely between the pretest and the first observation immediately after transfer to the two treatment units. Although these improvements were sustained over the 6-month treatment, no further improvement occurred later on. Data on frequency of care given by nursing staff indicated that there was an immediate significant decrease in care given in the general milieu program immediately after the programs began (sign test, $p < .022$), followed by a gradual increase over 6 months to the pretest levels.

Table 3 summarizes the stepwise linear discriminant analyses for the populations of

Table 2

Changes in Staff Observations From Before the Start of Programs to 6 Months After the Programs Began

Variable	Frequency ^a			n ^b
	Im- proved	No change	Worse	
Token economy				
Bizarre and unusual behaviors**	11	19	4	15
Conversations with other patients	16	8	16	32
Care given by nursing staff ^{*****}	16	16	2	18
Incontinence*	6	28	1	7
General milieu				
Bizarre and unusual behaviors***	11	24	1	12
Conversations with other patients	17	8	13	30
Care given by nursing staff	17	5	14	31
Incontinence ^{*****}	16	18	2	18

Note. Because all hypotheses were directional, one-tailed tests were used (by sign test, from Siegel, 1956, p. 250).

^a It was decided a priori that improvement would be indicated by increased frequencies of conversations with other patients, decreased amounts of care given by nursing staff, decreased frequencies of bizarre and unusual behaviors, and decreased incidences of incontinence.

^b In sign test.

* $p = .062$.

** $p = .059$.

*** $p = .003$.

**** $p < .001$.

Table 3

Multivariate Stepwise Discriminant Analyses for Improved and Unimproved Patients on Frequency of Bizarre and Unusual Behaviors for Token Economy and General Milieu Rehabilitation Programs

Variable	Token Economy		General milieu	
	M_{diff}	Order of entry	M_{diff}	Order of entry
Sex			.45385	2 (.5835)
Staff evaluations				
Dressing habits and abilities	.46640	6 (.3290)		
Physical condition	-.38735	4 (.4653)	9.86923	1 (.6597)
VIRO interaction				
Age				
Total years of hospitalization	4.09605	3 (.5601)		
PIP				
Anxious depression	-2.50197	5 (.4055)		
Care needed	.98419	2 (.6979)		
Psychotic disorganization	-8.16205	1 (.7960)		

Note. Linear discriminant coefficients appear in parentheses. VIRO = Vigor Intactness Relationship Orientation.

people who improved and did not improve on frequency of bizarre and unusual behaviors. The order of entry column indicates the rank order of the ability of the variables to distinguish between the two groups of patients when factors already entered in the discriminant function are controlled. Table 4 shows the results of multiple regression analyses on the same data, in which the order of entry of variables was the same as for the linear dis-

criminant analyses. Since the regression analyses were included in order to assess the amount of variance accounted for by each variable, the table shows the multiple correlations at each step, R^2 (amount of variance accounted for) at each step, and the simple correlation for each variable. Table 5 indicates the frequencies and percentages of misclassifications that would occur in these samples if the discriminant function were used to assign individuals to the improved or unimproved groups.

Table 4
Summary of Results from Multiple Regression Analysis With Order of Entry in Equation the Same as for the Discriminant Analyses

Variable	R	R^2	Simple r
Token economy program			
Psychotic disorganization	.199	.040	-.199
Care needed	.428	.183	.227
Total years hospitalization	.558	.312	.151
Physical condition	.609	.371	-.079
Anxious depression	.639	.408	-.163
Dressing habits	.646	.417	-.085
General milieu program			
VIRO interaction	.430	.184	.430
Sex	.632	.399	.554

Note. VIRO = Vigor Intactness Relationship Orientation.

In the general milieu program, populations that did and did not improve in bizarre and unusual behaviors could be discriminated best by the VIRO interaction score, followed by sex. Improved patients were generally less responsive on VIRO interaction ratings, and they tended more often to be males. On the token economy ward, PIP psychotic disorganization best discriminated between improved and unimproved patients, followed by PIP care needed, length of total hospitalization, staff evaluation of physical condition, PIP anxious depression, and staff evaluations of dressing habits. The improved patients generally had greater PIP psychotic disorganization, less PIP care needed, less total hospitalization, better staff ratings of physical condition, more anxious depression, and worse dressing habits.

Table 5
*Numbers of Cases Misclassified in Groups
 by the Stepwise Discriminant Analyses*

Group	Program	
	Token economy	General milieu
Improved as unimproved	1/11 (9.1)	4/11 (36.4)
Unimproved as improved	1/12 (4.3)	3/25 (12.0)
Total	2/34 (5.9)	7/36 (19.4)

Note. Numbers in parentheses are percentages.

Discussion

The token economy and general milieu programs were both successful in bringing about positive behavioral changes in a number of these chronic elderly mental hospital patients with organic diagnoses. Given the expectation that due to their diagnoses of irreversible chronic organic brain damage they should not be likely to improve, it is not unusual that the majority showed no change on the main outcome measure, frequency of bizarre and unusual behaviors. Most people showed no change, though significantly more people improved than declined. Still, the proportion of people who did improve (almost one third) supports current research in clinical gerontology, which indicates that many chronic elderly patients who were previously considered hopeless have rehabilitative potential. These findings are bolstered by the significant numbers who decreased in the frequency of incontinence, which had been considered by many staff members as a purely medical degenerative condition.

The lack of significant numbers of people increasing in their frequency of conversations with other patients is interesting. On both wards, although many conversed more frequently, almost equal numbers had less frequent conversations. I would like to believe that those who increased their conversations developed friendships and those who decreased their conversations stopped engaging in arguments or other undesirable behaviors. However, no data are available to explain this lack of findings, so one must assume that any

changes on this dimension are due to chance.

It is also of interest that care given by nursing staff for personal hygiene and self-care declined in the token economy program, but it declined at the start of the general milieu programs only to gradually increase to the pretreatment level. In the token economy it was obvious that the decline in staff care was a direct result of reinforcing patient behaviors, which lessened the need for staff help. In fact, nursing staff seemed particularly careful to reward behaviors that made their job easier. Many staff meetings were spent modifying the program so that other therapeutic behaviors would be the primary targets.

In the general milieu program, patients did more for themselves, which resulted in the initial drop in staff care. They required less help in basic care such as putting on one's clothes or taking a shower. However, as the program progressed, patients began to increase their demands on staff for additional forms of care and new services that were not given before. They demanded more meticulous grooming, such as careful hairstyling. In this instance the outcome variable that was defined a priori changed its meaning as the program progressed.

One possible explanation of these results is that the improvements were unrelated to the treatment methods—that resulted simply from a general "Hawthorne" effect of putting staff effort into patient care. If this explanation were correct, the results would indicate that any treatment is effective for this population when compared to no treatment at all. Clearly, this seems to be the case with regard to the decreases in the incidence of incontinence. These improvements occurred immediately on transfer to the new treatment environments but before the activities in the treatment programs had an opportunity to get under way.

An alternative hypothesis is that different types of patients are more likely to respond to each of the two programs. Support for this latter hypothesis can be seen in the results from the multiple discriminant analyses. Patients who improved in the token economy program could best be discriminated by their active psychotic disorganized behaviors, less

need for staff care, more recent admission to a mental hospital, better physical condition, active discontent, and sloppy dressing habits. This constellation of variables generally indicates the less "institutionalized" patients who are in better physical conditions and actively exhibit their troubles. However, in the general milieu program, patients who improved could be best discriminated on the basis of their lesser responsiveness to an interviewer.

These characterizations seem reasonable in view of the demands of each program. The token economy required patients to be sufficiently alert and motivated to engage in target behaviors that would result in reinforcement. Generally, more active involvement was necessary. However, the general milieu program did not demand active involvement, it only provided an enriched environment that was more supportive to individual needs.

One additional observation may be particularly relevant in deciding which type of program to conduct: Staff in the token economy program often reported that they were succeeding in their efforts, whereas staff in the general milieu program more often reported that they felt that they were getting nowhere. The reports of dismay in the milieu program occurred even when progress was being made. This difference in staff reactions may reflect the token economy staff's greater comfort with the highly structured token system. In the token economy, staff behaviors were clearly specified in most situations, and daily feedback was provided by charting of reinforcement frequencies for each patient. In the general milieu program, staff had to rely more on their intuition about how to behave. They were not given daily tallies that could provide feedback about the effectiveness of their actions.

Overall, this study suggests that different patient characteristics may discriminate between those who improve and those who do not in different types of rehabilitation programs. Further research in this area, and particularly cross-validations of these findings, may help clarify which patients benefit from which types of treatment. Perhaps further research will allow hospitals to rehabilitate a greater proportion of their patients by offering

alternative programs to meet the special needs of different patient populations.

References

- Atthowe, J. M., Jr., & Krasner, L. Preliminary report on the application of contingent reinforcement procedures (token economy) on a "chronic" psychiatric ward. *Journal of Abnormal Psychology*, 1968, 73, 37-43.
- Ayllon, T., & Azrin, N. *The token economy: A motivational system for therapy and rehabilitation of mental patients*. New York: Appleton-Century-Crofts, 1968.
- Ayllon, T., & Michel, J. The psychiatric nurse as a behavioral engineer. *Journal of Experimental Analysis of Behavior*, 1959, 2, 323-334.
- Bok, M. Some problems in milieu treatment of the chronic older mental patient. *Gerontologist*, 1971, 11, 141-147.
- Gottesman, L. E. Milieu treatment of the aged in institutions. *Gerontologist*, 1973, 13, 23-26.
- Gottesman, L. E., Quarterman, L. E., & Cohn, G. M. Psychological treatment of the aged. In C. Eisendorfer & M. P. Lawton (Eds.), *APA Task Force on aging report*. Washington, D.C.: American Psychological Association, 1973.
- Greenberg, D. J., Scott, S. B., Pisa, A., & Friesen, D. D. Beyond the token economy: A comparison of two contingency programs. *Journal of Consulting and Clinical Psychology*, 1975, 43, 498-503.
- Gripp, R. F., & Magaro, P. A. A token economy program evaluation with untreated control ward comparisons. *Behaviour Research and Therapy*, 1971, 9, 137-149.
- Gripp, R. F., & Magaro, P. A. The token economy program in the psychiatric hospital: A review and analysis. *Behaviour Research and Therapy*, 1974, 12, 205-228.
- Hoyer, W. J., Mishara, B. L., & Riedel, R. G. Problem behaviors as operants: Applications with elderly individuals. *The Gerontologist*, 1975, 15, 452-456.
- Kastenbaum, R., & Sherwood, S. VIRO: A scale for assessing the interview behavior of elderly people. In D. P. Kent, R. Kastenbaum, & S. Sherwood (Eds.), *Research planning and actions for the elderly*. New York: Behavioral Publications, 1972.
- Kleban, M. H., Lawton, M. P., Brody, E. M., & Moss, M. Behavioral observations of mentally impaired aged: Those who decline and those who do not. *Journal of Gerontology*, 1976, 31, 333-339.
- Krasner, L. The operant approach in behavior therapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change*. New York: Wiley, 1971.
- Lee, D. An adjunct to training psychiatric aides in behavior modification techniques. *Journal of Psychiatric Nursing and Mental Health Services*, 1969, 7, 169-171.
- Lorr, M., & Vestre, N. D. The Psychotic Inpatient Profile: A nurse's observation scale. *Journal of Clinical Psychology*, 1969, 25, 137-140.

- Maley, R. F., Feldman, G. L., & Ruskin, R. S. Evaluation of patient improvement in a token economy treatment program. *Journal of Abnormal Psychology*, 1973, 82, 141-144.
- Mueller, D. J., & Atlas, L. Resocialization of regressed elderly residents: A behavioral management approach. *Journal of Gerontology*, 1972, 27, 390-392.
- Shean, G. D., & Zeidberg, Z. Token reinforcement therapy: A comparison of matched groups. *Journal of Behavior Therapy and Experimental Psychiatry*, 1971, 2, 95-105.
- Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- Sherwood, S., Morris, J. N., & Barnhart, E. Developing a system for assigning individuals into an appropriate residential setting. *Journal of Gerontology*, 1975, 30, 331-342.

Received November 1, 1977 ■

Veridicality of Self-Report: Replicated Correlates of the Wiggins MMPI Content Scales

David Lachar and Richard S. Alexander
Lafayette Clinic, Detroit, Michigan, and Wayne State University

Correlates of the 13 Minnesota Multiphasic Personality Inventory Wiggins content scales were identified using both single-sample and cross-validation paradigms in a sample of 384 male clients who received a mental health evaluation at a military medical facility. Characteristics of both high-scoring and low-scoring clients were obtained for each scale, and interpretations of high scale elevations are proposed that reflected scale content, psychometric properties, and correlate characteristics. The 48 replicated and 28 single-sample correlates descriptive of high scorers significant at .01 reflected either the substantive nature of each scale or suggested primary characteristics expected of high scorers on each scale, whereas the additional 69 single-sample correlates significant at .05 often provided the descriptions needed to "round out" each scale interpretation. Evaluation of the correlates of low scorers essentially supported the position that low scores reflect the absence of descriptors characteristic of high elevations on the same scale. A factor analysis of the content, clinical profile, A, and R scales supported the interpretive intent of the content scales, as well as suggested their relative vulnerability to a defensive response set.

The 13 content scales of the Minnesota Multiphasic Personality Inventory (MMPI) were constructed by Wiggins (1966) to study the relation between item content and scale validity. They were developed by applying both psychometric and intuitive procedures to the original content classifications of Hathaway and McKinley (1951). Each scale item was assigned exclusively to its respective scale only if it obtained a point-biserial correlation with the total scale score in excess of .30, which also exceeded its correlation with the remaining content scales.

Initial study of the content scales revealed meaningful differences between various normal and psychiatric samples and between samples defined by traditional diagnostic classifications (e.g., schizophrenic psychoses, brain disorders, etc.). The factor structure of the scales was also supportive of their interpretive intent

(Wiggins, 1969). Additional construct validity has been demonstrated in a study that related content scale elevation to clinical profile code-type correlates (Payne & Wiggins, 1972), in studies that compared scale scores of groups differing in composition (Cohler, Grunebaum, Weiss, Hartman, & Gallant, 1975; Jarnecke & Chambers, 1977; Mezzich, Damarin, & Erickson, 1974; O'Neil, Teague, Lushene, & Davenport, 1975), as well as in studies that correlated content scale scores with other MMPI scales and other personality measures (Derogatis, Rickels, & Rock, 1976; Hoffmann & Jackson, 1976; Taylor, Ptacek, Carithers, Griffin, & Coyne, 1972; Wiggins, Goldberg, & Appelbaum, 1971).

Wiggins (1966) suggested that the content scales may serve as a supplementary source of information to interpretations derived from the empirically keyed MMPI profile scales. The content scales were constructed to measure the substance of the client's communication that is directed at the examiner. The interpretation of these scales represents a view of test response midway between the naive-rational and the radical-empirical perspec-

The authors wish to thank Kathleen Lennox and James L. Grisell for assistance in data analysis.

Requests for reprints should be sent to David Lachar, Division of Psychology, Lafayette Clinic, 951 East Lafayette, Detroit, Michigan 48207.

tives (Dahlstrom, 1969). Suggested interpretations of the content scales (Wiggins, 1966) reflect this interpretive intent by emphasizing that scale elevations reflect *admission* of various symptom clusters.

The present study seeks to evaluate the extent of agreement between self-report and clinical impression among clients by investigation of the external correlates of the Wiggins content scales. The study design and character of these scales also allows an examination of the relation between method of correlate selection and degree of conceptual fit with manifest scale content. A final goal of this investigation is the construction of interpretive statements for each scale to describe high-scoring clients.

Method

Subjects

The 384 male subjects were United States Air Force personnel and their dependents who had received a psychiatric evaluation between July 1972 and March 1973 at Wilford Hall Medical Center, Lackland Air Force Base, San Antonio, Texas. Mean age of this sample was 26.3 years (range = 18-57), and 95.4% were under 24 years of age. The mean educational level was 12.5 years, with 24% having more than a high school education. Marital status was about evenly divided between single (49%) and married (45%), and 6% were either divorced or widowed. These subjects were seen by different diagnostic components of the Department of Mental Health: 35% were inpatients on a psychiatric service, 18% were inpatients on general medical or neurology/neurosurgery services, 39% were seen at the outpatient clinic, and the remaining 8% were evaluated by the mental hygiene clinic during participation in basic training.

Measures

Each subject completed the standard booklet form of the MMPI as a routine part of his evaluation. These MMPIs were obtained as part of a previously reported study of rated accuracy of an automated interpretation system for the clinical profile scales (Lachar, 1974). *T* scores for the content scales were obtained using norms reported by Wiggins et al. (1971), and raw scores for profile scales and Scales *A* and *R* (Welsh, 1956) were converted to *T* scores using the Hathaway and Briggs (1957) norms.

Adjective correlates were collected from the clinician who had evaluated each subject using a mental health evaluation form, which was completed when the MMPI was administered. In addition to various identification, demographic, and clinical data, this

evaluation form permitted the description of each client by checking 81 possible descriptive adjectives. These adjectives were subjectively placed under the headings affect ($n = 18$), interpersonal relations ($n = 10$), motor behavior ($n = 9$), psychological efficiency ($n = 4$), patient-therapist relationship ($n = 3$), history ($n = 7$), thought process ($n = 8$), thought content ($n = 7$), and physical complaints ($n = 15$). Since many of the physical complaint descriptors were selected infrequently, these 15 adjectives were also collapsed into a composite variable labeled *physical complaint present*.

Procedure

MMPI protocols of dubious validity ($F > 25$ raw score) or substantial missing data ($Q > 30$ raw score) were omitted from this analysis, reducing the study sample by 21 subjects to 363. All subjects were then designated as either high scorers, low scorers, or "other" for each of the 13 content scales. Though score distributions allowed designation of $T < 41$ as the criterion inclusion rule for low scorers, the initial criterion inclusion rule of $T > 69$ for high scorers was modified to $T > 59$ for the Manifest Hostility, Authority Conflict, Feminine Interests, Hypomania, and Religious Fundamentalism scales to form sufficiently large criterion groups for these scales.

Correlates were selected for the total sample (criterion group vs. all remaining protocols) using the chi-square statistic at the .05 and .01 levels. In addition, each criterion group and the remaining sample were randomly split, and the adjective correlates significant at least at the .10 level in both samples were designated as replicated, with a resultant joint probability of at least .01.

To assist in definition of the construct dimensions related to each content scale, *T* values for the clinical profile scales, *A*, *R*, and the 13 content scales were intercorrelated. This correlation matrix was then submitted to a principal components factor analysis, and the resulting factor matrix was rotated to a varimax criterion.

Results

Criterion sample size, correlates replicated at .01, and correlates not replicated but significant at .01 or .05 for the total sample are presented below.¹ Forty-eight correlates replicated for high-elevation criterion groups, whereas only 18 correlates cross-validated in the analysis of low-score criterion groups. Approximately twice the number of correlates

¹ Adjective base rates and criterion group frequencies for significant correlates and scale intercorrelations are available on request from the first author.

were obtained by these three selection methods for the criterion groups defined by high elevations than for criterion groups defined by low elevations (145 vs. 88).

Organic Symptoms (ORG)

$T > 69$ ($n = 79$). Replicated at .01: impotent/decreased libido, fatigue, insomnia, poor memory, anorexia, headache, joint pains. Single sample significant at .01: depressed, difficulty in concentration, paucity of ideation, back pain, loss of consciousness, physical complaint present. Single sample significant at .05: less hyperactive/hypomanic, chest pain.

$T < 41$ ($n = 24$). Replicated at .01: less depressed. Single sample significant at .05: shallow affect, talkative, less fatigue.

Poor Health (HEA)

$T > 69$ ($n = 56$). Replicated at .01: chest pain. Single sample significant at .01: anorexia, diarrhea, shortness of breath. Single sample significant at .05: worrisome, constipation.

$T < 41$ ($n = 19$). Replicated at .01: less depressed, shallow affect. Single sample significant at .05: less anxious, less worrisome, less withdrawn.

Depression (DEP)

$T > 69$ ($n = 87$). Replicated at .01: impotent/decreased libido, retarded (motor), sense of inadequacy/inferiority. Single sample significant at .01: depressed, withdrawn. Single sample significant at .05: guilty, less inappropriate affect, combative when intoxicated, paucity of ideation.

$T < 41$ ($n = 32$). Replicated at .01: less depressed, less withdrawn, less sense of inadequacy/inferiority. Single sample significant at .05: less anxious, less guilty, less worrisome, talkative, less insomnia, less financial problems, less ideas of reference, physical complaint present.

Poor Morale (MOR)

$T > 69$ ($n = 65$). Replicated at .01: depressed, combative when intoxicated, paucity

of ideation, sense of inadequacy/inferiority, anorexia. Single sample significant at .01: fearful/phobic, tremulous. Single sample significant at .05: guilty, less inappropriate affect, less labile, suspicious, tearful, withdrawn, retarded (motor), insomnia, less religiosity.

$T < 41$ ($n = 64$). Replicated at .01: less depressed, less fearful/phobic. Single sample significant at .01: less sense of inadequacy/inferiority. Single sample significant at .05: less anxious, less guilty, less fearful, less worrisome, less impotent/decreased libido, less withdrawn, less agitated/restless, less indecisive, less difficulty in concentration, less insomnia, defensive, less ideas of reference, nausea/vomiting.

Social Maladjustment (SOC)

$T > 69$ ($n = 62$). Replicated at .01: apathetic, depressed, fearful/phobic, worrisome, withdrawn, compulsive, retarded (motor), sense of inadequacy/inferiority, suicidal thoughts. Single sample significant at .01: guilty, insomnia. Single sample significant at .05: fatigue, less poor judgment, constipation.

$T < 41$ ($n = 69$). Replicated at .01: less depressed, less fearful/phobic, less perplexed, less withdrawn. Single sample significant at .01: less poor memory, malingering. Single sample significant at .05: less apathetic, less guilty, less suspicious, less tearful, less worrisome, amoral, less dependent, less passive, less indecisive, less retarded (motor), less difficulty in concentration, less confused, less sense of inadequacy/inferiority, less suicidal thoughts, headaches.

Manifest Hostility (HOS)

$T > 59$ ($n = 24$). Replicated at .01: hostile, less suspicious, assaultive. Single sample significant at .01: chest pain. Single sample significant at .05: moody, amoral, impulsive, combative when intoxicated.

$T < 41$ ($n = 48$). Single sample significant at .05: less depressed, less moody, less impotent/decreased libido, less agitated/restless, less insomnia, less ideas of reference, less sense of inadequacy/inferiority.

Family Problems (FAM)

$T > 69$ ($n = 86$). Replicated at .01: assaultive, impotent/decreased libido, insomnia, drug usage, marital conflict. Single sample significant at .01: agitated/restless, impulsive. Single sample significant at .05: excitable, less passive, combative when intoxicated, sense of inadequacy/inferiority, unrealistic feelings, less loss of consciousness.

$T < 41$ ($n = 34$). Replicated at .01: less depressed, homicidal, less marital conflict. Single sample significant at .01: abdominal pain. Single sample significant at .05: inappropriate affect, less sense of inadequacy/inferiority.

Authority Conflict (AUT)

$T > 59$ ($n = 94$). Replicated at .01: malingering. Single sample significant at .01: marital conflict, less autistic thought, less perfectionistic. Single sample significant at .05: assaultive, less withdrawn, destructive gestures, combative when intoxicated.

$T < 41$ ($n = 49$). Single sample significant at .01: less immature. Single sample significant at .05: less moody.

Feminine Interests (FEM)

$T > 59$ ($n = 75$). Replicated at .01: perplexed, difficulty in concentration, unrealistic feelings. Single sample significant at .01: suicide attempts, religiosity. Single sample significant at .05: homosexual, less impulsive, indecisive, insomnia, less malingering, less combative when intoxicated, confused, delusions, hallucinations.

$T < 41$ ($n = 39$). Single sample significant at .05: agitated/restless, malingering.

Phobias (PHO)

$T > 69$ ($n = 47$). Replicated at .01: fearful/phobic, worrisome, retarded (motor), anorexia. Single sample significant at .01: depressed. Single sample significant at .05: anxious, withdrawn, tremulous, poor memory, paucity of ideation, chest pain, joint pains.

$T < 41$ ($n = 42$). Single sample significant at .01: physical complaint present. Single

sample significant at .05: abdominal pain, back pain, visual problems.

Psychoticism (PSY)

$T > 69$ ($n = 54$). Replicated at .01: retarded (motor), autistic thought, paucity of ideation. Single sample significant at .01: anorexia. Single sample significant at .05: perplexed, suspicious, worrisome, less dependent, impotent/decreased libido, disorganized thought, incoherent, less perfectionistic, hallucinations, ideas of reference.

$T < 41$ ($n = 21$). Single sample significant at .05: less suspicious, less insomnia.

Hypomania (HYP)

$T > 59$ ($n = 104$). Replicated at .01: excitable, immature, hyperactive/hypomanic. Single sample significant at .01: agitated/restless. Single sample significant at .05: less compulsive, destructive gestures, less retarded (motor), malingering, less suicidal thoughts.

$T < 41$ ($n = 28$). Single sample significant at .01: less immature. Single sample significant at .05: nausea/vomiting.

Religious Fundamentalism (REL)

$T > 59$ ($n = 40$). Replicated at .01: less alcohol excess. Single sample significant at .01: delusions, religiosity. Single sample significant at .05: less impulsive, less drug usage, less marital conflict, autistic.

$T < 41$ ($n = 84$). Replicated at .01: homicidal, impulsive, drug usage. Single sample significant at .05: less moody, destructive gestures, confused.

The results of the factor analysis are presented in Table 1. Five factors were obtained that accounted for 96% of the common variance among these 28 scales. The first factor accounted for 57.2% of the variance and reflected somatic complaints (ORG, HEA, H_s , H_y) and a secondary emphasis of psychological discomfort (D , Pt , Sc). The second factor accounted for 16.3% of the variance and appeared to be organized around informant response style. Measures of symptom denial (L , K , R) characterized one end of this dimension, whereas the other was defined by ad-

Table 1

Rotated Factor Matrix of the Content, Validity, Clinical, A, and R Scales

Scale	1	2	3	4	5	h^2
Wiggins content						
ORG	.72*					.76
HEA	.75					.76
DEP		-.42	.73			.88
MOR		-.40	.76			.86
SOC			.91			.84
HOS		-.77				.73
FAM		-.42			.41	.49
AUT		-.73				.59
FEM				.78		.60
PHO			.54			.48
PSY		-.59				.74
HYP		-.80				.68
REL						.09
Clinical profile						
L		.54				.37
F		-.45			.51	.74
K		.61				.83
Hs	.93					.87
D	.49		.65			.75
Hy	.85					.79
Pd					.66	.59
Mf				.72		.60
Pa					.40	.61
Pt	.48		.55		.42	.77
Sc	.50		.43		.50	.81
Ma		-.63				.58
Si			.95			.93
Welsh factor						
A		-.47	.75			.92
R		.66				.54
% of variance	57.2	16.3	11.1	7.0	4.4	

Note. ORG = Organic Symptoms; HEA = Poor Health; DEP = Depression; MOR = Poor Morale; SOC = Social Maladjustment; HOS = Manifest Hostility; FAM = Family Problems; AUT = Authority Conflict; FEM = Feminine Interests; PHO = Phobias; PSY = Psychoticism; HYP = Hypomania; REL = Religious Fundamentalism.

*Factor loadings less than .40 have been omitted.

mission of pathology (*F*, *A*). Although only 1 of the 10 clinical profile scales loaded on Factor 2, 7 of the content scales obtained such loadings. The especially high and primary loadings of scales HOS, AUT, PSY, and HYP on Factor 2 suggested their susceptibility to a defensive response set. The third factor accounted for 11.1% of the variance and clearly represented a dimension characterized by depression (DEP, *D*), anxiety (PHO, *Pt*, *A*), and social withdrawal and alienation (SOC, *Si*, *Sc*). The fourth factor accounted for 7% of the variance and uniquely represented traditional sex role interest pattern (FEM, *Mf*). The final factor represented 4.4% of the vari-

ance and was characterized by Scales FAM, *F*, *Pd*, *Pa*, *Pt*, and *Sc*. This factor appeared to reflect both poor interpersonal relationships and the causal and resultant traits associated with interpersonal conflict.

Proposed interpretations of high content scale scores are presented in the appendix. These interpretations reflect a combination of scale item content, empirically supported scale correlates, and their frequency, as well as scale intercorrelations.

Discussion

The correlates obtained for the Wiggins content scales demonstrated substantial agree-

ment between client self-report and clinician evaluation. Evaluation of correlates of high scale elevations by method of correlate selection suggested that both the 48 replicated and 28 nonreplicated correlates selected at .01 directly reflected either the substantive nature of their corresponding content scale or primary characteristics expected of clients who score highly on each scale. This substantive agreement raises doubts as to the frequently stated need for the cross-validation of correlates (Boerger, Graham, & Lilly, 1974; Gynther, Altman, & Sletten, 1973; Lewandowski & Graham, 1972; Marks, Seeman, & Haller, 1974) if the level of significance for correlate selection is sufficiently conservative in analysis of sufficiently large samples. Inspection of the additional 69 correlates selected for the total sample at .05 reveals that they often provide the descriptors needed to "round out" each scale interpretation. The correlates of HOS ($T > 59$), for example, include "hostile, less suspicious, assaultive, chest pain" at .01, whereas inclusion of .05 descriptors adds "moody, amoral, impulsive, and combative when intoxicated."

Evaluation of the 25 correlates selected at .01 and the 69 correlates selected at .05 for low-elevation criterion samples supports the position that low scores reflect the absence of descriptors characteristic of high elevations on the same scale. This analysis does not exclude the possibility that some of the content scales reflect bipolar dimensions, as the potential correlate pool for this study did not include any favorably worded descriptors. An artifact of this analysis appears to be several physical complaint correlates of low elevations [e.g., "abdominal pain" for PHO ($T < 41$) and FAM ($T < 41$)] resulting from sample inclusion of physically ill patients referred for evaluation who evidenced no symptoms of psychological disturbance.

The susceptibility of personality scales with high content saturation to response biases such as defensiveness is illustrated by previously demonstrated negative relationships between subtlety of scale item content and the amount of K correction applied to the 10 clinical profile scales of the MMPI.

The second factor obtained in the current factor analysis highlighted the susceptibility of the majority of the content scales to such a defensive response set. A correlate of MOR ($T < 41$), "defensive," also suggested that lowered content scale scores are likely to result from intentional client distortion. Clinicians applying the content scales should be aware that high elevations represent significant probability correlates, whereas low elevations are less likely to suggest the absence of these symptoms—especially if Scales L or K obtain some elevation.

In spite of our hope, inspired by adherence to dustbowl empiricism, of discovering a number of unexpected correlates, few were obtained. One interesting finding was that REL, and to some extent FEM, elevations represent inhibition of acting out in male psychiatric populations. That is, the admission of religious beliefs is strongly associated with decreased substance abuse, marital conflict, and impulsive behavior, whereas low REL elevations relate to the replicated correlates "homicidal, impulsive, drug usage." Also, it appears that PSY ($T > 69$) in a predominantly nonpsychotic sample, reflects not only low base-rate psychotic behaviors, but it also suggests perplexity, suspiciousness, and worry.

References

- Boerger, A. R., Graham, J. R., & Lilly, R. S. Behavioral correlates of single-scale MMPI code-types. *Journal of Consulting and Clinical Psychology*, 1974, 42, 398-402.
- Cohler, B. J., Grunebaum, H. U., Weiss, J. L., Hartman, C. R., & Gallant, D. H. Perceived life-stress and psychopathology among mothers of young children. *American Journal of Orthopsychiatry*, 1975, 45, 58-73.
- Dahlstrom, E. G. Recurrent issues in the development of the MMPI. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications*. New York: McGraw-Hill, 1969.
- Derogatis, L. R., Rickels, D., & Rock, A. F. The SCL-90 and the MMPI: A step in the validation of a new self-report scale. *British Journal of Psychiatry*, 1976, 128, 280-289.
- Gynther, M. D., Altman, H., & Sletten, I. W. Replicated correlates of MMPI two-point code-types: The Missouri actuarial system. *Journal of Clinical Psychology Monographs*, 1973, 29, 263-289.
- Hathaway, S. R., & Briggs, P. F. Some normative data on new MMPI scales. *Journal of Clinical Psychology*, 1957, 13, 264-268.

- Hathaway, S. R., & McKinley, J. C. *The Minnesota Multiphasic Personality Inventory manual* (Rev. ed.). New York: Psychological Corporation, 1951.
- Hoffmann, H., & Jackson, D. N. Substantive dimensions of psychopathology derived from MMPI content scales and the Differential Personality Inventory. *Journal of Consulting and Clinical Psychology*, 1976, 44, 862.
- Jarnecke, R. W., & Chambers, E. D. MMPI content scales: Dimensional structure, construct validity, and interpretive norms in a psychiatric population. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1126-1131.
- Lachar, D. Accuracy and generalizability of an automated MMPI interpretation system. *Journal of Consulting and Clinical Psychology*, 1974, 42, 267-273.
- Lewandowski, D., & Graham, J. R. Empirical correlates of frequently occurring two-point MMPI code-types: A replicated study. *Journal of Consulting and Clinical Psychology*, 1972, 39, 467-472.
- Marks, P. A., Seeman, W., & Haller, D. L. *The actuarial use of the MMPI with adolescents and adults*. Baltimore: Williams & Wilkins, 1974.
- Mezzich, J. E., Damarin, F. L., & Erickson, J. R. Comparative validity of strategies and indices for differential diagnosis of depressive states from other psychiatric conditions using the MMPI. *Journal of Consulting and Clinical Psychology*, 1974, 42, 691-698.
- O'Neil, H. F., Teague, M., Lushene, R. E., & Davenport, S. Personality characteristics of women's liberation activists as measured by the MMPI. *Psychological Reports*, 1975, 37, 355-361.
- Payne, F. E., & Wiggins, J. S. MMPI profile types and self-report of psychiatric patients. *Journal of Abnormal Psychology*, 1972, 79, 1-8.
- Taylor, J. B., Ptacek, M., Carithers, M., Griffin, C., & Coyne, L. Rating scales as measures of clinical judgment. III: Judgments of the self on personality inventory scales and direct ratings. *Educational and Psychological Measurement*, 1972, 32, 543-557.
- Welsh, G. S. Factor dimensions A and R. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: University of Minnesota Press, 1956.
- Wiggins, J. S. Substantive dimensions of self-report in the MMPI item pool. *Psychological Monographs*, 1966, 80(22, Whole No. 630).
- Wiggins, J. S. Content dimensions in the MMPI. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications*. New York: McGraw-Hill, 1969.
- Wiggins, J. S., Goldberg, L. R., & Appelbaum, M. MMPI content scales: Interpretive norms and correlations with other scales. *Journal of Consulting and Clinical Psychology*, 1971, 37, 403-410.

Appendix

Proposed Interpretation of High MMPI Content Scale Scores

Organic Symptoms (T > 69)

This individual has admitted to a variety of sensory, motor, or general somatic concerns that may be related to psychological discomfort and general malaise as well as to reduced effectiveness in completing daily tasks. Clients who obtain high Organic Symptoms elevations may complain of lack of stamina and strength and may present physical symptoms that often indicate emotional conflict, such as problematic headache or back pain.

of interest in the environment, pessimism, self-criticism, and brooding. In client populations, social withdrawal, a negative self-concept, guilt feelings, and a reduced activity level may be suggested.

Poor Morale (T > 69)

Inventory responses reflect a pervasive lack of confidence in one's abilities and a history of failure, which is related to these perceived limitations. Clients who obtain high Poor Morale elevations may be insecure, despondent, withdrawn, intropunitive, and oversensitive, and may become easily upset by the actions of others.

Social Maladjustment (T > 69)

Endorsed item content reflects a lack of social skill and poise, discomfort in social interaction, and resultant inhibition and social isolation. In client populations, this lack of social supports may be associated with a negative self-image, feelings of despair or fearfulness, thoughts of suicide, or a defensive orientation characterized

Poor Health (T > 69)

A significant number of physical complaints are reflected by item endorsement centering mainly around the digestive system. Individuals who obtain high Poor Health elevations are often considerably worried about their health. Cardiac and pulmonary complaints are also occasionally reported.

Depression (T > 69)

This individual has admitted to symptoms associated with problematic depression, such as lack

by apathy and limited activity or compulsive attention to detail.

Manifest Hostility (T > 59)

This individual admits to problems in adjustment related to unmodulated expression of anger, resentment of perceived injustices, need for interpersonal dominance, and limited self-control. In client populations, the combination of hostility, moodiness, and impulsivity may be associated with assaultive or other antisocial or violent behavior.

Family Problems (T > 69)

Inventory responses include admission of pathology in and among family members. A history of poor relationships with parents is suggested, as well as the absence of positive supports in current family interactions, whether with parents, spouse, or extended family. Patient male: In adult male clients, admission of family pathology may reflect not only marital conflict but may also suggest intolerant, overreactive individuals and a negative self-concept. Drug abuse and other destructive behavior may be associated.

Authority Conflict (T > 59)

Endorsed item content reflects the belief that interpersonal relations are often exploitive in nature. Disregard for principles of ethical conduct and truthfulness is suggested, as well as a tendency to minimize the negative impact of antisocial behavior. In client populations, these attitudes may be associated with problematic overassertive and manipulative social relations. Conflict with relatives may result.

Feminine Interests (T > 59)

Inventory responses suggest an interest in pursuits traditionally labeled as feminine and/or dislike of activities stereotyped as masculine. Patient male: In male clients, this interest pattern may be associated with an indecisive, passive orientation that has proven to be problematic. Conflict may lead to confusion or self-

blame. Evaluation for suicidal ideation or previous attempts is suggested.

Phobias (T > 69)

This individual admits to a variety of fears and appears to be significantly uncomfortable in many situations. Clients who obtain high PHO elevations are viewed as more anxious, tremulous, worrisome, and phobic than most patients. Depression and social withdrawal may also be indicated.

Psychoticism (T > 69)

Inventory responses include admission of unusual experiences and beliefs, many of which may include a clearly paranoid component. In client populations, this response pattern often suggests an individual who finds comprehension of human motives and behavior difficult and is consequently suspicious of and worried about others. Symptoms associated with a psychotic adjustment, such as ideas of reference, hallucinations, and autistic or disorganized thought, may be present.

Hypomania (T > 59)

This individual's self-description suggests a fast personal tempo characterized by enthusiasm, cheerfulness, and perhaps irritability or emotional lability. Clients who obtain high Hypomania elevation are often described as immature, hyperactive, excitable, agitated, and restless. They are unlikely to respond intropunitively to conflict and may manipulate others to reach their goals.

Religious Fundamentalism (T > 59)

Endorsed item content reflects strong religious beliefs and religiously motivated behavior. In client populations, this orientation suggests a reduced probability of substance abuse, impulsive behaviors, and conflict with family members. Expression of strong religious beliefs may, at times, reflect a delusional system and associated thought disorder.

Received November 1, 1977 ■

Effects of Film Modeling on the Reduction of Anxiety-Related Behaviors in Individuals Varying in Level of Previous Experience in the Stress Situation

Barbara G. Melamed
University of Florida

Richard Yurcheson, E. Louis Fleece, Steven Hutcherson, and
Roland Hawes
Case Western Reserve University

The influence of film preparation on 80 children undergoing three dental sessions (prophylaxis, dental examination, and dental restorative treatment) was evaluated with respect to (a) peer modeling versus demonstration of procedures and (b) amount of information. It was found by evaluating self-report and behavioral and visceral-arousal indices in a 2×2 factorial design that children exposed to a peer-model videotape presentation immediately preceding their own restorative treatment exhibited fewer disruptive behaviors and reported less apprehension than those watching a videotaped demonstration without a peer model. The modeling film elicited less heart rate activity in the subjects than the demonstration. The younger children (4-6 years) had lower self-reports of fear after viewing a more complete synopsis of what to expect, whereas the older children (8-11 years) had the lowest report of fears after viewing the peer model receiving a local anesthetic and brief intraoral examination. Children with previous treatment experience benefitted most from viewing the peer model undergoing the entire restorative procedure or a demonstration of the administration of local anesthetic in the absence of a peer model. Children with no prior experience were sensitized by being shown this demonstration. Thus, it was concluded that the age and previous experience of the viewer were important factors in determining childrens' fear-related behaviors after exposure to preparatory stimuli.

A large literature indicates that modeling, either live or symbolic (filmed), is effective in reducing fearful avoidance behaviors and in increasing adaptive behaviors in a wide variety of situations (Bandura, 1969; Bandura & Menlove, 1968). The effectiveness of modeling cannot be assumed without understanding what information needs to be presented, how this information can best get across to

the observer, and how the previous experience of the individual in the situation modifies these considerations.

The current investigation addressed these issues in children facing a real-life stress: dental treatment. Early attempts to introduce children to dental treatment by use of modeling films were successful (Adelson & Goldfried, 1970; Machen & Johnson, 1974). Melamed and her colleagues (Melamed, Hawes, Heiby, & Glick, 1975; Melamed, Weinstein, Hawes, & Katin-Borland, 1975) demonstrated reduced disruptive behavior and lower ratings of anxiety in children who had had no previous dental treatment experience after the children viewed a cooperative peer model as compared with children in a

This study was supported by National Institute of Dental Research Grant DE-04243-02.

Appreciation is expressed to Carol Gee and to the Pediatric Dental Clinic Auxiliary Staff of Rainbow Babies and Childrens Hospital, Cleveland, Ohio.

Requests for reprints should be sent to Barbara G. Melamed, who is now at the Dept. of Psychology, Brogden Psychology Bldg., 1202 W. Johnson St., University of Wisconsin, Madison, WI 53706.

control condition who observed either no film or a film unrelated to dentistry.

However, when film modeling was compared with systematic desensitization or a placebo (warm interaction with the hygienist), its effectiveness became less certain. Machen and Johnson (1974) and Sawtell, Simon, and Simeonsson (1974) found contradictory results. Much of the discrepancy between the findings of different investigators was most likely due to a lack of attention to the subject's prior experience with dental treatment and the presentation of information in the modeling situation.

This study compared peer modeling with a demonstration of the same dental personnel and procedures in the absence of a child model. Both of these videotapes provided the subjects with information about impending procedures, but in addition the peer model provided a sample of adaptive behaviors that should be emitted during treatment.

The second variable, amount of information presented, was used to evaluate whether the child observer needs to be exposed to each step of the procedure or whether exposure to the most feared event, the anesthetic injection, would facilitate modeling. Kleinknecht, Klepac, and Alexander (1973) suggested that dental fears can best be conceptualized as learned responses to painful stimuli during dental treatment.

The evaluation of the effectiveness of film modeling in relation to previous experience of the observer allowed us to look at the influence of information presented on avoidance behaviors that were generated by vicarious learning (no actual prior experience) or by some combination of classical, operant, and vicarious conditioning.

Method

Subjects

Eighty children between the ages of 4 and 11 were selected from the pedodontic clinic of University Hospitals, Cleveland, Ohio. Mentally handicapped or physically disabled children were excluded. Children who needed or had received previous extractions were excluded. Our sample consisted of 58 black children and 22 white children of homogeneous lower socioeconomic status. Assignment was made to one of five film groups with balancing for age, sex, race, previous dental experience, and initial report of fear on the Children's Fear Survey Schedule (Scherer & Nakamura, 1968) modified for dental-

related fears. There were 48 children who had had one or more cavities restored and 32 without prior treatment experience.

Procedure

Subjects were observed during two clinic visits. At the first visit the hygienist performed a standard prophylaxis (Session 1). Immediately afterward, a dentist examined the child's mouth and prepared bitewing radiographs (Session 2). The same dentist restored one or two cavities in a 30-minute appointment 7-10 days later (Session 3). The experimental manipulation involved exposure of children separately to one of five videotapes according to group assignment immediately prior to their restorative treatment session. Dentists were matched for experience, and both they and the independent observers were unaware of group assignment.

Groups

Long model. A 7-year-old black child is viewed undergoing a dental restorative treatment procedure. This 10-minute tape included the examination, injection, cavity preparation, and placement of the restoration. The boy remains cooperative and fearless throughout. The dentist and assistant are neutral. They instruct the model but do not use positive or negative reinforcement.

Long demonstration (demo). The 10-minute videotape was matched for the auditory tract, with the same dentist and assistant demonstrating the identical procedures without a child model in the chair. The dentist does say what behavior he expects in the child.

Short model. The same child is shown receiving the anesthetic injection followed by an oral examination. He remains cooperative throughout. This videotape runs approximately 4 minutes.

Short demonstration (demo). The dentist and assistant demonstrate the anesthetic injection and the oral examination without a child in the chair. This videotape runs approximately 4 minutes.

Unrelated control film. A videotape of *Evan's Corner* (Bostustow, 1969) pictures a 7-year-old black boy fixing a corner of the living room as his special place. This served as a control for subjects' exposure to a film.

Assessment battery. Previous research reporting low intercorrelations between different anxiety measures (Lang, 1968, 1978; Venham, Bengtson, & Cipes, 1977; Bernstein & Kleinknecht, Note 1) led us to examine different aspects of a child's subjective, behavioral, and physiological anxiety. Table 1 indicates the variety of measures used and the times at which they were assessed. On the first visit each mother (or guardian) completed a questionnaire about the child's behavior problems (Peterson, 1961) and one concerning her anxiety about dentistry and how she perceived her youngster would respond to dental treatment (Johnson & Baldwin, 1969). The Children's Fear Survey Schedule (CFSS; including 15 specific dental items) was administered orally to the child at this time and was repeated immediately prior to treatment

Table 1
Summary of Times of Measurement

Measure	Session 1 dental hygiene			Session 2 dental exam			Behavior manipulation			Session 3 dental treatment	
	Pre	During	Post	Pre	During	Post	Pre	During	Post	During	Post
MAQ	X										
BPCL	X										
CFSS	X								X		
FT	X					X	X		X		X
PSI	X			X		X	X		X		X
BPRS		X			X					X	
GSR								X		X	
HR								X		X	
DRA			X			X					X
DRC			X			X					X
ORA			X			X					X
ORC			X			X					X

Note. MAQ = maternal anxiety questionnaire; BPCL = Behavior Problem Checklist; CFSS = Children's Fear Survey Schedule; FT = fear thermometers; PSI = Palmar Sweat Index; BPRS = Behavior Profile Rating Scale; GSR = galvanic skin response; HR = heart rate; DRA = dental rating of anxiety; DRC = dental rating of cooperation; ORA = observer rating of anxiety; ORC = observer rating of cooperation.

after viewing the appropriate videotape. The children rated their degree of fear (on a 5-point fear thermometer) prior to the hygienist's treatment, following the dentist's examination, before and after the film, and following the restorative session. The Palmar Sweat Index (PSI; Thomson & Sutarman, 1953) was obtained at these same times. Observer ratings of anxiety were obtained during each session by an independent observer on the Behavior Profile Rating Scale (BPRS), a quantified measure of disruptive behavior. This instrument has demonstrated reliability and validity (Klorman, Ratner, King, & Sven, 1977; Melamed, Hawes, Heiby, & Glick, 1975; Melamed, Weinstein, Hawes, & Katin-Borland, 1975.) In addition, the dentist and the observer rated each subject on a 10-point scale of cooperation (1 = good, 10 = poor) and anxiety (1 = low to 10 = high).

Galvanic skin response and cardiac responses were measured during film viewing and treatment by means of a Grass polygraph. Data were stored on a Vetter FM tape recorder, and the heart rate data were available for analysis by means of a PDP-12 computer program. Five 30-sec samples from each film were selected for analysis. To compare videotapes of differing lengths, the total time of each videotape was divided into sixths. The five samples analyzed were taken at the first, second, third, fourth, and fifth one sixth of the videotapes starting with the third sample, which was chosen as the exact temporal middle of all films. Thus, the first two samples and the last two samples were the same proportional distance from the beginning and the end of each film. Heart rate difference scores were computed using median heart period in

a one-min prefilm resting baseline and subtracting from that the median heart rate period in milliseconds for each of the five samples for each individual.

In addition, the median heart rate for 10-sec intervals preinjection scene, during the injection, and postinjection were sampled. Difference scores were calculated by subtracting each from the preinjection value. These heart rate difference scores were analyzed to assess increments and decrements in heart rate in response to the specific content of the local anesthetic injection, regardless of the time into the videotape.

Results

There were no significant pretreatment group differences, except on the PSI. Each dependent variable was analyzed in a 2×2 factorial design to evaluate the effect of modeling versus demonstration, long versus short, and all possible interactions involving these factors and the repeated measure of time of assessment. Separate analyses were repeated with age, sex, race, and experience as separate main effects. Age was defined in three age blocks: 4-6 years; 6-8 years; and 8-11 years. Experience was dichotomized as no previous treatment versus previous restorations. A one-way analysis of variance was computed on all dependent measures to de-

termine whether an unrelated film control group, run after the main experiment, was significantly different from any of the treatment groups.¹ Appropriate subsequent *t* tests were used. All *p* values are for two-tailed tests.² An intercorrelation matrix of all measures at all assessment periods was obtained.

Modeling Versus Demonstration

The greater effectiveness of the modeling over the demonstration videotape was substantiated across most measures. Children viewing the modeling tapes reported fewer overall fears than children viewing the demonstration videotapes, $F(1, 56) = 4.26$, $p < .04$. The race of the subject did affect the subjective report. White children reported fewer fears than black children after viewing the modeling tape, $t(29) = 2.13$, $p < .05$, despite the fact that the child model was black. In fact, the black children reported fewer fears on dental items than whites after seeing the demonstration tapes, $t(31) = 2.6$, $p < .02$.

The behavioral measures were concordant with self-report findings. There was a borderline significance ($p < .06$), with the children viewing the modeling tape exhibiting less disruptive behavior than those viewing the demonstration of the same procedures. This interaction with sessions achieved significance, $F(1, 56) = 7.06$, $p < .01$, as illustrated in Figure 1. Immediately after viewing the peer model, children were less disruptive during

actual restorative treatment than those shown the demonstration version.

The type of videotape observed also yielded significant differences in the observer's ratings of anxiety and cooperation when age was taken into account. The 6- to 8-year-old children were rated as less anxious if they had seen the modeling film than if they had viewed the demonstration, $t(23) = 2.11$, $p < .05$, as illustrated in Figure 2. The same age group was also rated as more cooperative during restorative treatment if they had had the opportunity to view the peer model than if they had viewed the demonstration version for both observer ratings, $t(23) = 2.3$, $p < .05$, and dentist ratings, $t(23) = 3.3$, $p < .01$.

The Palmar Sweat Index data analyzed by covariance analysis showed a significant increase prefilm to postfilm, regardless of which videotape was seen. There was a significant correlation of the initial PSI with children's general fears ($r = .30$, $p < .02$) and specific dental fears ($r = .31$, $p < .01$), as reported on the Children's Fear Survey Schedule administered immediately preceding restorative treatment. The initial scores on the CFSS also correlated positively with the PSI taken immediately before the videotape was shown ($r = .31$, $p < .01$).

The heart rate data for the injection scenes also reflected a difference due to the type of film seen, $F(1, 51) = 4.69$, $p < .035$. There was a greater increase in children's heart

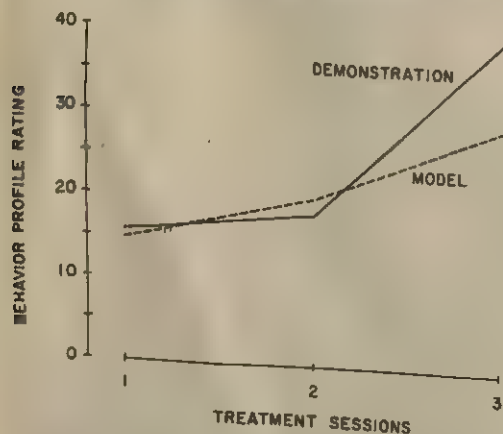
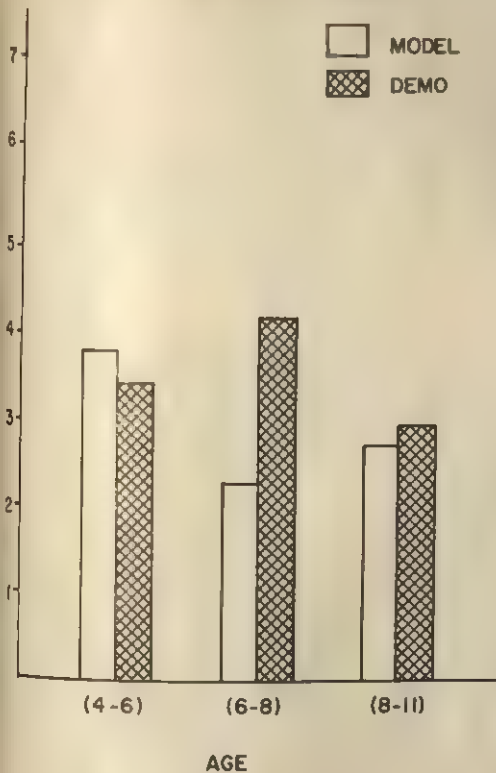


Figure 1. Degree of disruptiveness during dental treatment and type of videotape preparation.

¹ Because the subjects in the unrelated film group were all run after the completion of the videotape conditions, the results are suspect. The failure to replicate high levels of disruptive behavior found in our previous studies (Melamed, Hawes, Heiby, & Glick, 1975; Melamed, Weinstein, Hawes, & Katin-Borland, 1975) may be accounted for by the fact that all of these subjects were hooked up to a polygraph during treatment, thus limiting movement and likelihood of disruption. Therefore, only significant differences between the treatments and this condition were reported in this study. Appropriate replication is needed prior to interpreting the lack of group differences when one-way analyses of variance, including the control group, were performed.

² Appendix A, which has a table of all comparisons made, is available on request from the first author.

OBSERVER'S RATING OF ANXIETY



AGE

Figure 2. Observer's rating of children's anxiety by age and type of videotape preparation.

rates from the preinjection period to the during- and after-injection segments of the demonstration videotapes than in the heart rates of children who viewed the modeling versions. In fact, the group means in heart rate difference scores for the modeling conditions ($M = 8.69$ for during and $M = 11.55$ for postinjection) indicated a further decrease in heart rate as compared with the preinjection values.

Amount of Information

Although a main effect for length of the film was only apparent in the heart rate data, the self-report data did reflect an effect of film length when the age of the child was considered. No differences were found for the amount of information on any of the observational measures.

Children in the youngest age group, 4-6 years, reported fewer general fears after seeing the longer version as compared with the

short videotapes, $t(16) = 2.35$, $p < .05$. However, it should be noted that there was a main effect of age on the fear thermometer, another self-report measure. The youngest children gave higher ratings of self-reported fear than other age groups regardless of the type of film, $F(2, 65) = 4.95$, $p < .01$. The oldest children, 8-11 years, had the lowest report of dental-related fears after viewing

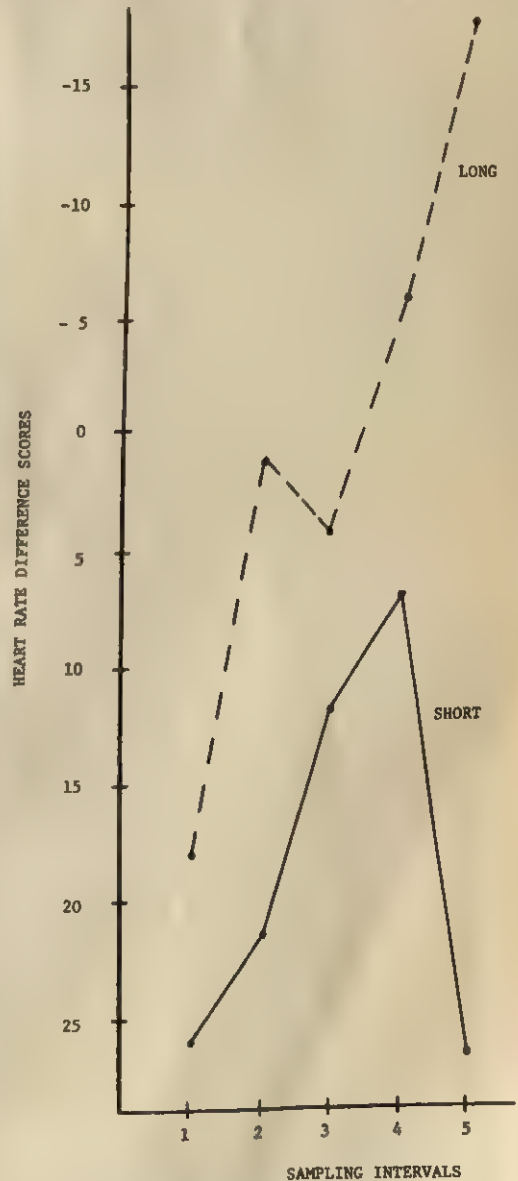


Figure 3. Heart rate difference scores for long and short videotapes across the five sampling intervals.

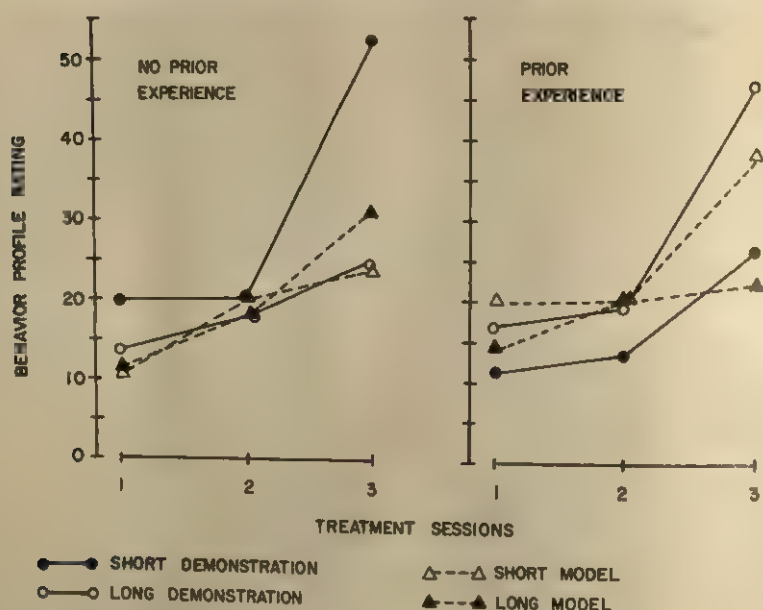


Figure 4. Degree of disruptiveness during dental treatment as effected by previous experience and type of videotape preparation.

the short model as compared with the short demonstration, $t(8) = 2.44$, $p < .05$.

Heart rate data did yield a significant effect of length of the film, $F(4, 204) = 3.07$, $p < .02$, with long films eliciting higher heart rate difference scores. In the interaction between length and interval, $F(4, 212) = 2.64$, $p < .04$, shown in Figure 3, it is clear that there exists a linear positive relationship between heart rate acceleration and time into videotape for the longer, more informative videotape. The children viewing the shorter tapes on the other hand showed an initial peak during the second, third, and fourth sampling points (film content overlaps in part with the injection segment) and a return to the initial heart rate difference scores by the end of the videotapes. In fact, the heart rate periods in response to the short versions never exceeded the prefilm baseline values, whereas the last two sampling points on the long version (film content shows restorative procedures after injection) exceeded the baseline values in the direction of increased heart rate (decreased heart period).

Previous Experience

The effectiveness of the different types of preparatory films on the behavior of the child

was influenced by previous experience, $F(1, 56) = 7.06$, $p < .01$. The difference caused by previous experience was not noted in any of the self-report measures.

Group differences are illustrated in Figure 4. The Behavior Profile Rating scores of Session 2 were used as a covariate in determining the significant differences between film types in groups defined as no prior experience versus previous experience because of the high correlations between these scores and the BPRS during Session 3 ($r = .38$, $p < .001$). The children with no prior experience showed significantly less disruptive behavior after viewing a short model or a long demonstration prior to their first restorative treatment, as compared with viewing the short demonstration ($p < .01$). In fact, the short-demonstration group showed higher degrees of disruption when compared with children who viewed the unrelated control film ($p < .02$). The children with previous treatment showed less disruptive behavior after viewing the long modeling film as compared with the long demonstration, $t(14) = -2.75$, $p < .007$. However, there were no significant differences between the long model and other videotapes, including the unrelated film. Overall, the dentists rated the

inexperienced children as decreasing in cooperation from the dental examination to the dental treatment, $t(48) = 2.0$, $p < .05$, whereas the change in cooperation for experienced subjects was insignificant. The correlations between BPRS scores for Session 2 and Session 3 were highly positive for experienced subjects ($r = .44$, $p < .002$), as were the dentists' ratings of their anxiety between Session 1 and Session 3 ($r = .63$, $p < .001$).

Intercorrelation Matrices

The data revealed good test-retest reliability within measures. Table 2 shows the high positive correlations that resulted between comparisons on the self-report measures. Table 3 indicates that the degree of disruptiveness (BPRS) during the first two sessions was highly related to the children's behavior during restorative sessions across all subjects. It was interesting to note that the greatest concordance between dentist and observer ratings of anxiety and cooperation occurred during the actual restorative treatment session. Table 4, which illustrates the comparison between measures tapping different systems (i.e., self-report and behavioral), is not unexpected in its sparse number of significant correlations. The most noteworthy is the fact that the children's self-reported dental fears just prior to the restoration correlated significantly with the degree

of disruptiveness during that session ($r = .29$, $p < .01$). This measure of anticipatory anxiety also correlated with the observer and dentist ratings of anxiety and cooperation during the treatment session.

The Table 5 measures of age, maternal anxiety, and general problem behaviors (BPCL) yielded some interesting correlations with the other dependent variables. Age correlated consistently with the child's self-report of anxiety, with younger children reporting more anxiety. Dentists and observers also tended to rate the younger child as more anxious. It is interesting that there was no significant correlation between the Behavior Problem Checklist and any of the BPRS measures. This would support the contention that this scale is measuring something other than just the general level of disruptiveness, hyperactivity, or management problems of the children. The relationship reported earlier in the literature between maternal anxiety and children's anxiety is not supported. In fact, the mothers' concern about their children is negatively associated with the children's self-report of anxiety prior to actual dental treatment.

Discussion and Conclusions

Peer Modeling Versus Demonstration

Does peer modeling reduce anxiety more than a demonstration of the same procedure?

Table 2
Pearson Product-Moment Correlations Between Self-Report Measures of Anxiety

Variable	2	3	4	5	6	7	8	9
1. CFSSA (1)	.71***	.82***	.59***	.26*	.20	.23*	.18	.30**
2. CFSSA (2)		.68***	.87***	.27*	.14	.32**	.41***	.30**
3. CFSSB (1)			.72***	.35***	.02	.28*	.31**	.23*
4. CFSSB (2)				.22*	.09	.33**	.55***	.33**
5. FT (1)					.21	.32**	.14	.26*
6. FT (2)						.15	.15	.33**
7. FT (3)							.44***	.19
8. FT (4)								.22*
9. FT (5)								

Note. $N = 80$. CFSSA = Children's Fear Survey Schedule, full scale score; CFSSB = Children's Fear Survey Schedule, dental items only; FT = fear thermometer. Arabic numerals in parentheses indicate measurement time.

* $p < .05$.
 ** $p < .01$.
 *** $p < .001$.

Table 3
Pearson Product-Moment Correlations Between Observational Ratings of Anxiety and Cooperation

	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. BPRS (1)	.38***	.40***	.33**	.32**	.45***	.46***	.21	.19	.57***	.27*	.45***	.71***	.39***	.32**
2. BPRS (2)		.35***	.06	.19	.23*	—	.17	.17	.05	.13	.13	.21	.27*	.26*
3. BPRS (3)			.12	.19	.68***	.16	.11	.62***	.11	.24*	.72***	.23*	.25*	.75***
4. DRA (1)				.38***	.27*	.78***	.17	.20	.57***	.50***	.25*	.56***	.39***	.33*
5. DRA (2)					.42***	.35***	.51***	.23*	.45***	.53***	.21	.38***	.24*	.20
6. DRA (3)						.33**	.25*	.75***	.34**	.39***	.62***	.40***	.25*	.61***
7. DRC (1)							.15	.16	.62***	.44***	.30**	.70***	.36***	.22*
8. DRC (2)								.32**	.32**	.25*	.09	.25*	.22*	.20
9. DRC (3)									.10	.11	.50***	.20	.14	.64***
10. ORA (1)										.59***	.30**	.67***	.22*	.11
11. ORA (2)											.24*	.37***	.31**	.13
12. ORA (3)												.41***	.24*	.67***
13. ORC (1)													.57***	.23*
14. ORC (2)														—
15. ORC (3)														

Note. *N* = 80. BPRS = Behavior Profile Rating Scale; DRA = dental rating of anxiety; DRC = dental rating of cooperation. ORA = observer rating of anxiety; ORC = observer rating of cooperation. Arabic numerals in parentheses indicate measurement time (Sessions 1-3).

* *p* < .05.
** *p* < .01.
*** *p* < .001.

Table 4
Pearson Product-Moment Correlations Between Self-Report and Observational Measures

	CFSSA (1)	CFSSA (2)	CFSSB (1)	CFSSB (2)	FT (1)	FT (2)	FT (3)	FT (4)	FT (5)
BPRS (1)	.11	.15	.25*	.19	.09	-.08	.11	.15	.01
BPRS (2)	.16	.18	.11	.11	-.05	.14	.28*	-.03	.09
BPRS (3)	.13	.20	.17	.29**	-.16	-.03	.25*	.19	.00
DRA (1)	.27*	.31**	.37***	.31**	.28*	-.09	-.09	.05	.01
DRA (2)	.18	.19	.27*	.23*	.17	.04	.11	.23*	.12
DRA (3)	.12	.20	.18	.24*	-.05	-.09	.19	.17	-.03
DRC (1)	.15	.29**	.34**	.32**	.31**	-.13	.06	.12	.01
DRC (2)	-.04	-.11	-.02	-.09	.00	.31**	.03	-.03	-.07
DRC (3)	.12	.15	.08	.18	-.04	.04	.14	.20	-.09
ORA (1)	.04	.18	.20	.18	.30**	.04	-.03	.02	.00
ORA (2)	-.02	.15	.17	.19	.13	-.07	.11	.04	-.05
ORA (3)	.17	.18	.21	.25*	-.05	-.03	.09	.13	.13
ORC (1)	.18	.28*	.33**	.28*	.14	-.06	.05	.20	.02
ORC (2)	.05	.07	.17	.13	-.15	.05	.02	.15	-.05
ORC (3)	.04	.12	.09	.20	-.16	-.05	.04	.04	.00

Note. $N = 80$. BPRS = Behavior Profile Rating Scale; DRA = dental rating of anxiety; DRC = dental rating of cooperation; ORA = observer rating of anxiety; ORC = observer rating of cooperation. CFSSA = Children's Fear Survey Schedule, full scale score; CFSSB = Children's Fear Survey Schedule, dental items only; FT = fear thermometer. Arabic numerals in parentheses indicate measurement time.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

dures? There is evidence that this is true. Children who observed a peer-model film reported less dental and general fears than those who observed the demonstration film. In fact, the observational data are congruent with this in that children observing the peer model cooperated with the dentist and exhibited fewer disruptive behaviors. The children in the 6- to 8-year-old group were rated as less anxious and more cooperative if they had seen the peer-model film as opposed to a demonstration. Dentists also rated these children as more cooperative. These data support the general finding that children closest in age to the filmed model were the most susceptible. (It is interesting that black children did not reduce their self-reported fear after seeing a black child model as much as white children did.) The physiological data are also congruent. Higher sympathetic arousal (heart rate) was produced in response to the demonstration tape, whereas deceleration occurred in response to peer modeling. Thus, modeling films that show peers of similar age cooperating with dentists have a more favorable effect on the self-reported apprehension, actual behaviors of the

observing children, and autonomic indices of observers than do mere desensitization or exposure to impending events.

Amount of Previous Experience

The length of the film affected the self-reported apprehension, but not the behavior, of the children. The youngest children, ages 4-6, had the lowest reports of fear with longer versions of the film regardless of type of presentation (model vs. demo). The more information imparted, the more the heart rate difference scores increased. Therefore, it is not sufficient to give more information, as this can have a sensitizing effect, if the format of the presentation (model or demo) is ignored.

Effects of Previous Experience

The effects of having had prior dental experience were seen in terms of the children's behavior during dental restorative treatments. The dentists rated the children with no prior experience as being less cooperative during treatment than during the dental examination. The child who had al-

Table 5
Pearson Product-Moment Correlations Among Age, Maternal Anxiety Questionnaire (MAQ) and Behavior Problem Checklist (BPCL) Scores, Measures of Self-Report, and Observational Ratings

	Age	MAQ	BPCL
Age		.00	-.12
CFSSA (1)	-.07	-.19	.14
CFSSA (2)	-.28*	-.33**	.34**
CFSSB (1)	-.11	-.23*	.14
CFSSB (2)	-.25*	-.26*	.28*
FT (1)	-.29**	-.15	-.09
FT (2)	-.25*	-.03	-.13
FT (3)	-.15	-.07	.11
FT (4)	-.16	-.07	.27*
FT (5)	-.16	.15	-.04
MAQ	.00		-.34**
BPCL	-.12	-.34**	
BPRS (1)	-.14	.10	-.11
BPRS (2)	-.07	.14	.07
BPRS (3)	-.03	.04	.22
DRA (1)	-.20	-.28*	.22*
DRA (2)	-.22*	.05	.07
DRA (3)	-.10	.03	.21
DRC (1)	-.15	-.19	.15
DRC (2)	-.19	.00	.00
DRC (3)	-.08	-.08	.33**
ORA (1)	-.27*	-.11	.01
ORA (2)	-.14	-.01	.04
ORA (3)	.06	.01	.21
ORAC (1)	-.17	.01	.01
ORC (2)	-.13	.04	.01
ORC (3)	-.05	.11	.23*

Note. $N = 80$. BPCL is a measure of the mother's rating of child's general behavior problems; CFSSA = Children's Fear Survey Schedule, full scale score; CFSSB = Children's Fear Survey Schedule, dental items only; FT = fear thermometer; BPRS = Behavior Profile Rating Scale; DRA = dental rating of anxiety; DRC = dental rating of cooperation; ORA = observer rating of anxiety; ORC = observer rating of cooperation. Arabic numerals in parentheses indicate measurement time.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

ready been to a dentist did not get significantly more anxious or disruptive during treatment. The effectiveness of different types of preparatory films on reducing disruptive behavior was related to previous experience. Children with no prior experience did best when given prior exposure to either a long demonstration tape or a short version with a peer model. They were sensitized by seeing a

short demonstration. On the other hand, children who have already experienced dental restorations behave much better if they view a long modeling version. Thus, when these children received maximum information regarding what to expect and how to behave, they were most cooperative during treatment. However, the lack of difference between the long model and the short demonstration at tests to the ability of some children to use their past experience in knowing how to behave, even in the absence of a model.

This study has raised important issues regarding the blanket assumption that filmed modeling reduces fear-related behaviors. It was found that there is a greater reduction of both self-reported and actual fear behavior during treatment when information is presented through a peer model. Whether this was due to greater information regarding behavioral expectation or whether viewing an identification figure increased modeling cannot be definitively stated. The physiological data would support that peer modeling produced a catharsis for model condition, since observers heart rates decreased with exposure.

The importance of age and previous experience in choosing appropriate preparatory material was borne out. The child most similar in age to the peer model being portrayed benefitted the most in his or her degree of cooperation. Race had a paradoxical effect, with white children showing lower self-reports of fear than blacks after seeing a black model.

In terms of previous experience, some important issues regarding the influence of prior knowledge on selection of type and format of information were raised. If the child has never actually undergone dental treatment but has formed a conception through vicarious processes, such as being told that a shot will be given, reminding him or her of this event in the absence of a model showing how to handle it (short demonstration) will sensitize him or her and increase the degree of disruption. On the other hand, the more experienced child will benefit most from being prepared for this noxious event briefly or by

seeing another child going through each step of the procedure in a cooperative manner.

Thus, the dental setting allows the researcher to investigate many questions regarding prevention and modification of stress responses, taking into account developmental and learning differences in information-processing abilities of the coping individuals.

Reference Note

1. Bernstein, D., & Kleinknecht, R. *Assessment of fear of dentistry*. Paper presented at the 10th annual convention of the Association of the Advancement of Behavior Therapy, Washington, D.C., December 1976.

References

- Adelson, R., & Goldfried, M. Modeling and the fearful child patient. *Journal of Dentistry for Children*, 1970, 37, 476-489.
- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Bandura, A., & Menlove, F. Factors determining vicarious extinction of avoidance behavior through symbolic modeling. *Journal of Personality and Social Psychology*, 1968, 8, 99-108.
- Bostustow, S. (Producer). *Evan's Corner*. Santa Monica: Bailey Film Associates Educational Media, 1969.
- Johnson, R., & Baldwin, D., Jr. Maternal anxiety and child behavior. *Journal of Dentistry for Children*, 1969, 36, 87.
- Kleinknecht, R., Klepac, R., & Alexander, L. Origins and characteristics of fear of dentistry. *Journal of the American Dental Association*, 1973, 86, 842-848.
- Klorman, R., Ratner, J., King, J. Jr., & Sveen, O. Pedodontic patients' uncooperativeness and maternal anxiety. *Journal of Dental Research*, 1977, 56, 432. (Abstract)
- Lang, P. J. Fear reduction and fear behavior: Problems in treating a construct. In J. M. Schlien (Ed.), *Research in psychotherapy* (Vol. 3). Washington, D.C.: American Psychological Association, 1968.
- Lang, P. J. Anxiety: Toward a psychophysiological definition. In H. Akiskal & W. Webb (Eds.), *Psychiatric diagnosis: Exploration of biological predictors*. New York: Spectrum, 1978.
- Machen, J., & Johnson, R. Desensitization, model learning, and the dental behavior of children. *Journal of Dental Research*, 1974, 53, 83-89.
- Melamed, B. G., Hawes, R., Heiby, E., & Glick, J. The use of film modeling to reduce uncooperative behavior of children during dental treatment. *Journal of Dental Research*, 1975, 54, 797-801.
- Melamed, B. G., Weinstein, D., Hawes, R., & Katin-Borland, M. Reduction of fear-related dental management using filmed modeling. *Journal of the American Dental Association*, 1975, 90, 822-826.
- Peterson, D. Behavior problems of middle childhood. *Journal of Consulting Psychology*, 1961, 25, 205-209.
- Sawtell, R., Simon, J., & Simeonsson, R. Effects of five preparatory methods upon child behavior during the first dental visit. *Journal of Dentistry for Children*, 1974, 37-45.
- Scherer, M. W., & Nakamura, C. Y. A fear survey schedule for children (FSS-FC): A factor analytic comparison with manifest anxiety (CMAS). *Behaviour Research and Therapy*, 1968, 173-182.
- Thomson, M. L., & Sutarman, M. The identification and enumeration of active sweat glands in man from plastic impressions of the skin. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 1953, 47, 412-417.
- Venham, L., Bengston, D., & Cipes, M. Children's response to sequential dental visits. *Journal of Dental Research*, 1977, 56, 454-459.

Received November 2, 1977 ■

A and B Undergraduate Interviewers of Schizophrenic and Neurotic Inpatients: A Test of the Interaction Hypothesis

Daniel F. Barnes
Counseling Center
Loyola University of Chicago

Juris I. Berzins
University of Kentucky

In an attempt to elucidate further the personological basis of the differential compatibility of A and B therapists with schizophrenic and neurotic patients, this study required A and B undergraduate volunteers (20 males, 20 females) to conduct 20-minute interviews with male state hospital inpatients (40 schizophrenics, 40 neurotics) in a 2 (interviewer A-B status) \times 2 (interviewer sex) \times 2 (patient type) factorial design. As expected from studies of the personality correlates of A-B status, many more B than A interviewers "looked forward" to conducting the interviews. Once in the interview situation, however, A-type interviewers elicited better self-disclosure from schizophrenic patients than did Bs, whereas the latter outperformed As with neurotic patients. The results are discussed in terms of a personological formulation that considers interviewer effectiveness to be a joint function of interviewer personality characteristics and the situational context.

The A-B variable emerged from a series of studies conducted by Whitehorn and Betz (1954) in the early 1950s. In these studies, therapists (arbitrarily labeled A and B) were found to be differentially effective with schizophrenic patients (As more effective than Bs). However, with neurotic patients, McNair, Callahan, and Lorr (1962) found Bs to be more effective than As. These findings suggested an "interaction hypothesis": As are more effective with schizophrenics than are Bs, whereas the latter are more effective with neurotics than are As. Subsequent analogue and clinical studies have generally supported the interaction hypothesis (e.g., Berzins, Ross, & Friedman, 1972; King & Blaney, 1977; Matthews & Burkhart, 1977).

This article is based on a study presented by the first author under the supervision of the second author as partial fulfillment of the research requirements leading to the PhD at the University of Kentucky.

The authors wish to thank Marian Grace, William A. Hunt, and Ronald E. Walker for their ideas and comments.

Requests for reprints should be sent to Daniel F. Barnes, Counseling Center, Loyola University of Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626.

Several explanatory formulations of the A-B interaction effects (e.g., the "complementarity" hypothesis, perceptual-cognitive style compatibility) have been proposed. One such formulation can be called the personological hypothesis, which evolved from a study by Berzins, Barnes, Cohen, and Ross (1971), who showed that A-B differences can be explained in personality terms. Using the Personality Research Form (PRF, Jackson, 1967), they found that A-type persons were characterized by cautious self-expression, social ineptness, and a restricted cognitive scope, whereas B subjects appeared socially ascendant and "open" to complex experiences. The single most differentiating personality dimension was "harmavoidance" (As > Bs), denoting cautious avoidance of risk taking.

Berzins, Dove, and Ross (1972) cross-validated the Berzins et al. (1971) study with very large samples of college males and females, male therapists, and male college clinic patients. In spite of striking intergroup differences in PRF scale scores, A and B persons, across samples, showed distinctive personality profiles. In terms of five "core" scales, B-type persons were less harmavoid-

ant but more dominant, variety seeking, sentient, and "counterdependent" than were A-type persons.

The personality correlates of A-B status, in our view, should be regarded as depicting the *everyday* behaviors of A and B persons. Asking A and B college students to interview a schizophrenic state hospital inpatient, however, does not require the student to display everyday, accustomed behaviors. Rather, he or she is required to interact in a presumably "helpful" fashion with someone who is severely maladjusted. These situational demands, furthermore, need not be experienced similarly by As and Bs because of their differing levels of adaptation, self-confidence, and personal adjustment. Therefore, one would expect A- and B-type persons to display their "everyday" behaviors *before* they conduct an interview with a "state hospital patient" but not while interviewing. Accordingly, the following hypotheses were advanced.

1. Prior to meeting any patient, B-type persons should look forward to interviewing state hospital patients to a greater extent than should A-type persons.

2. Once in the situation, A-type persons should elicit better self-disclosure from schizophrenic patients and display greater non-verbal immediacy with them, whereas B-type persons should perform similarly with neurotic patients than should be the case under the opposite or "incompatible" pairing conditions.

3. After the interview, more positive reactions about the interview should be elicited from members of optimal pairings, as compared with members of nonoptimal pairings, as a consequence of greater interview success.

Method

Interviewers

A total of 131 student volunteers (56 males, 75 females), enrolled in the summer school session in psychology, filled out the PRF (Jackson, 1967) and the A-B scale (19 scored items). From this group, 40 students (the first 10 male and female As and the first 10 male and female Bs) were selected as interviewers. Using the Berzins, Dove, and Ross (1972) normative data, A status was defined by

scores between 0 and 7 for males and between 0 and 4 for females; B status, by scores 13 to 19 for males and 10 to 19 for females.

Patients

Schizophrenic ($n=40$) and neurotic ($n=40$) male inpatients at the Eastern State Hospital (Lexington, Kentucky) were selected on the basis of symptom clusters rather than formal diagnosis. Phillips and Rabinovitch (1958) distinguished three symptom cluster categories: (a) self-deprivation, turning against the self (TAS); (b) avoidance of others (AVOS); and (c) self-indulgence, turning against others (TAO). The first two categories, TAS and AVOS, were considered prototypic of neurotic and schizophrenic status, respectively. Using a procedure developed by Green (1971) and Welch (1971), a patient was classified as AVOS if the number of AVOS symptoms noted was equal to at least two and if the number of TAS plus TAO symptoms was less than the number of AVOS symptoms. Similarly, a patient was classified as TAS if he had at least two TAS symptoms and fewer (than TAS) AVOS plus TAO symptoms noted in the clinical folder.

Procedure

Each interviewer was to meet one AVOS and one TAS patient, with each interview lasting 20 minutes. The interviews were held in a moderately large one-way vision room that contained no furniture except for 15 chairs located on the perimeter of the room.

Before the first interview, each interviewer was instructed to "interview each patient as completely as you can by getting as much information as you can." To help the interviewer gain some degree of comfort and structure, the experimenter provided a deck of 12 cards, each with a possible interview topic on it. Of the 12 topics, 6 had been preterated as personal in content (e.g., "What aspects of your personality do you dislike?"), and 6 were neutral in content (e.g., "What is your favorite reading matter?"). After the interviewers had familiarized themselves with the topics (no directions were given to restrict the interview to these topics, however), each interviewer was asked casually if he/she was "looking forward to the interview," and was introduced to the patient and left to choose seats and conduct the interview. Following the first interview, the patient and the interviewer were asked (independently) how he or she felt about the interview. The same procedure was followed with the second patient. Both interviews were terminated by the experimenter's entering the room.

Dependent Measures

During the interviews, two advanced graduate students (one clinical, one nonclinical) observed

the interaction through the one-way mirror and rated selected aspects of the participants' behavior.

The main dependent measure was depth of patient self-disclosure, rated by both judges on a 0-3 scale (Berzins, Ross, & Cohen, 1970). Depth of self-disclosure ratings were assigned to every interview topic perceived as distinct by each judge; the topics were later subgrouped according to the personal-neutral distinction. In addition, at the 1st, 5th, 10th, 15th, and 20th minutes of the interview, the judges used multipoint scales to rate the interviewers' nonverbal behavior in 11 categories. These were taken from Mehrabian (1968) and included measures of "immediacy" or liking (physical proximity, eye contact, orientation of torso), relaxation or status (asymmetrical placement of arms and legs, sideways lean, reclining position, hands relaxed), and activity (facial expressiveness, rate and flow of speech).

The experimenter (the first author) also unobtrusively recorded interviewers' responses to the preinterview question ("Are you looking forward to the interview?"), and, after each interview, he asked the participants independently (a) "Did you like the interview?" (b) "Did you like the person you talked to?" and (c) "Throughout the interview, were you basically tense or relaxed?" Responses to these questions were coded as 0 or 1. Since the experimenter could not help being aware of the group membership of some patients and some interviewers, he attempted to ask all questions as impartially and equivalently as possible.

Design

The experimental design was a 2 (interviewer A-B status) \times 2 (interviewer sex) \times 2 (patient type), with the order of presentation of the last factor counterbalanced to control for order effects. There were 10 dyads per cell, with each interviewer having seen one AVOS and one TAS patient. Of the 40 interviewers, 20 saw the AVOS patient first, and 20 saw the TAS patient first.

Results

A correlational analysis was conducted to determine whether the A-B scale correlates in the PRF, originally demonstrated on much larger samples (Berzins, Dove, & Ross, 1972), were replicable within the summer school classes from which the interviewers were drawn. As expected, the A-B scale scores related negatively to the Harmavoidance and Succorance scales but positively to the Sentience, Dominance, and Change scales of the PRF, thereby replicating prior results.

Interjudge Reliability

Since the judges had strong agreement regarding the number of different topics covered by each interviewer, $r(78) = .93$, $p < .0001$, only those topics that were rated by both judges were considered. For individual topics ($n = 1,118$ across the 80 dyads), the interjudge coefficient for patient self-disclosure ratings was .69 ($p < .0001$), but the coefficient for mean self-disclosure per dyad, $r(78) = .89$, $p < .0001$, indicated that the judges used the self-disclosure rating scale in a consensual manner. Of the 11 nonverbal measures of interviewer behavior, three (eye contact, verbal fluency, and rate of speech) were not rated reliably (r s less than .27), but the remainder showed interjudge coefficients ranging from .64 to .98 (all p s $< .0001$). Only the latter were analyzed.

Preinterview Measures

We had hypothesized that prior to meeting any state hospital inpatient, the B-type interviewers would look forward to the interviews to a greater extent than A-type interviewers. This hypothesis was supported by the data, $\chi^2(1) = 16.94$, $p < .001$, with the effect especially pronounced among male interviewers, $\chi^2(1) = 16.36$, $p < .001$.

Intrainterview Measures

A four-factorial analysis of variance (incorporating the patient sequence as a separate factor) conducted across all intrainterview measures showed eight significant interaction effects, seven of which involved the patient sequence variable. That is, A-type interviewers in the AVOS-TAS sequence (AVOS patient seen first) and Bs in the TAS-AVOS sequence obtained more self-disclosure from patients than did interviewers paired oppositely. In general, although results with the "first patient seen" conformed to the interaction hypothesis, these trends also transferred themselves to the second interview, even though another (compatible or incompatible) patient now comprised the stimulus. These order effects replicate those of Berzins and Seidman (1968, 1969).

Since the interviewer's performance with the "first" patient cannot be affected by the impending second interview, analyzing the first interviews should afford a clearer appraisal of the interaction hypothesis in this experiment, although at a cost of statistical power (cf. Berzins & Seidman, 1968, 1969). Table 1 presents the results that were significant in three-factual (interviewer A-B status, interviewer sex, patient type) analyses of variance of all intrainterview measures.

Of the four significant patient-type main effects, three involved the self-disclosure variable. The TAS patients clearly were more self-disclosing than were AVOS patients, with regard to the personal, neutral, and total topics discussed within the interview. (The total topic category also includes the mean self-disclosures on topics introduced by interviewers, i.e., topics not on the cards.)¹ Since the AVOS patients may be regarded as more regressed than TAS patients, this result is not at all surprising and replicates the results of Green (1971) and Welch (1971) with Veterans Administration patients.

The other main effect involving the AVOS-TAS distinction (asymmetrical placement of arms by the interviewer) indicates that AVOS patients elicited higher levels of this index of relaxation (Mehrabian, 1968) than did TAS patients. This finding seems consistent with the notion that more schizophrenics exude lower status than do neurotics.

The main effects involving interviewer sex concerned one measure of nonverbal immediacy (proximity) and one measure of activity (facial expressiveness). Males sat closer to the patients than did females, but females were markedly more expressive than males. These results might best be understood in terms of sex role stereotypic behaviors.

The finding that A-type interviewers sat closer to the patients than did Bs is consistent with the results of the Berzins et al. (1970) study and of the Green (1971) and Welch (1971) studies, in which the A interviewer was rated as "warmer" than the B interviewer. Note, however, that this index of behavioral "approach" strikingly contra-

Table 1

Significant Main and Two-way Interaction Effects in the Three-Factorial Analyses of Variance of Intrainterview Measures

Variable	Type		F(1, 2)
Main effects			
Patient type	AVOS	TAS	
Self-disclosure, total	2.03	2.33	7.28**
Self-disclosure, personal topics	1.81	2.27	10.76***
Self-disclosure, neutral topics	1.96	2.28	8.88***
Asymmetry of arms	1.77	.80	5.00*
Interviewer sex	M	F	
Physical proximity	6.50	4.35	5.12*
Facial expressiveness	1.54	3.75	25.40***
Interviewer A-B status	A	B	
Physical proximity	6.60	4.25	6.12**
Interviewer A-B Status × Patient Type Interaction			
	AVOS	TAS	
Self-disclosure			
Total			
A	2.12	2.19	4.23*
B	1.95	2.48	
Personal topics			
A	2.05	2.12	7.91***
B	1.56	2.42	
Neutral topics			
A	2.10	2.11	8.16***
B	1.83	2.44	

Note. M = male, F = female; AVOS = avoidant of others (schizophrenic prototype), TAS = turning against self (neurotic prototype).

* $p < .05$.

** $p < .02$.

*** $p < .01$.

dicts expectations based on these interviewers' responses to the preinterview question, "Are you looking forward to the interview?"

Turning now to the two-way interaction effects in Table 1, it is clear that the main effects denoting the greater self-disclosure of TAS patients is qualified prominently by interviewers' A-B status. That is, A-type interviewers obtained equivalent self-disclosures from both types of patients, whereas Bs

¹ There were no main or interaction effects involving interviewer A-B status for the number of personal, neutral, extraneous, or total topics discussed. The average number of topics per dyad were 4.3 personal, 4.6 neutral, and 4.9 extraneous.

obtained considerably higher self-disclosures from TAS than AVOS patients. (Orthogonal comparisons between patient types within Bs were significant at $p < .005$ for total, personal, and neutral topics.) However, on personal and neutral (but not total) topics, A-type interviewers obtained significantly higher self-disclosures from AVOS than TAS patients (personal topics, $p < .01$; neutral topics, $p < .05$). Also, with TAS patients, B-type interviewers outperformed As on total ($p < .05$) and neutral ($p < .02$) but not on personal topics. Taken together, these results offer substantial support to the interaction hypothesis and are also consistent with the findings of Berzins et al. (1970) with addict patients, and those of Greene (1971) and Welch (1971) with Veterans Administration inpatients.

Two additional interaction effects (Interviewer A-B Status \times Interviewer Sex \times Patient Type), not shown in Table 1, involved one measure of nonverbal immediacy (orientation of torso) and one measure of relaxation (asymmetry of arms). Close examination of these effects, considered jointly, suggest that under compatible pairing conditions, male interviewers manifested greater immediacy but less relaxation than did male interviewers paired incompatibly. This attitude of poised attention appears consistent with the self-disclosure data and the interaction hypothesis. Female interviewers, however, did not follow this pattern; rather, they appeared to exhibit greater immediacy and relaxation with the AVOS patient than with the TAS patient. These differences invite clarification in further research.

Postinterview Measures

Our third hypothesis stated that members of compatibly paired dyads would show more positive postinterview reactions than members of incompatibly paired dyads. Since the variability of participants' reactions was constrained by ceiling effects (e.g., in only eight instances did an interviewer report having disliked the patient), the hypothesis was not supported statistically. However, when negative reactions were given, they

were consistent with our hypothesis and with the intrainterview self-disclosure data.

Discussion

The preinterview data clearly supported the first hypothesis and also offered behavioral validation of the interviewer's everyday personality characteristics, as reflected by PRF scale scores. However, since responses to the "looking forward to" question were solicited by the experimenter, these differences, their magnitude notwithstanding, require replication with better control over possible experimenter effects.

Before discussing the intrainterview results, the patient sequence factor bears examination. In the present study as in earlier analogues (Berzins & Seidman, 1968, 1969), undergraduate As and Bs paired compatibly and incompatibly "transferred" their differential performances with the first patient to the second patient. This "transfer" of performance differences associated with initial pairings to later ones, contrary to Chartier (1971), in no way cancels out the initial differences. Rather, these sequence effects, demonstrated in both analogue and in vivo settings, deserve recognition in the design of further research.

Turning now to the interviews themselves, the degree of support accorded the interaction hypothesis by the self-disclosure data seems important particularly because this study paired untrained, unsophisticated interviewers with markedly disturbed inpatients in a realistic interview situation. In spite of this arrangement, one that should work against ready, let alone differential, elicitation of self-disclosures, the A-type interviewers were able to elicit as much self-disclosure from AVOS (schizophrenic prototype) as TAS (neurotic prototype) patients, and they outperformed Bs with the AVOS patients.

Recalling that both in terms of PRF scores and responses to the looking forward question, A-type interviewers appeared more reticent than did Bs before encountering any patient, the finding that once in the interview, they in fact sat closer to their patients

than did Bs is intriguing also. Since increased physical proximity seems an unlikely correlate of cautious and reticent attitudes, it seems plausible that the A-type interviewer's "transformation" may have been eventuated by some aspect of the actual interview situation.

Our personological formulation assumes that the interviewer's behavior (effective or ineffective) is a function of the interviewer's own everyday personality characteristics and the situational context, the latter comprised at least of the demand that one function as a "helping agent," and the perceived characteristics of the patient, for example, degree of psychopathology. In everyday life, the A-type interviewer indeed may be cautious and submissive, and the B-type interviewer may be risk oriented and dominant. Upon encountering a markedly disturbed patient, however, the cautious and submissive A-type interviewer may perceive the regressed patient as someone whom he/she can help. In this sense, the regressed patient "liberates" A-type interviewers to behave effectively, whereas less disturbed patients may render these interviewers uncertain about the outcomes and may "inhibit" their effectiveness. On the other hand, with less disturbed patients, the dominant, risk-taking, self-assured B interviewer may perceive the situation as a reasonable challenge; contrariwise, the more disturbed patient may be experienced by the B-type interviewer as boring or even hopeless. The results of this study generally support this formulation.

References

- Berzins, J. I., Barnes, D. F., Cohen, D. I., & Ross, W. F. Reappraisal of the A-B therapist "type" distinction in terms of the Personality Research Form. *Journal of Consulting and Clinical Psychology*, 1971, 36, 360-369.
- Berzins, J. I., Dove, J. L., & Ross, W. F. Cross-validation studies of the personality correlates of the A-B therapist "type" distinction among professionals and nonprofessionals. *Journal of Consulting and Clinical Psychology*, 1972, 39, 388-395.
- Berzins, J. I., Ross, W. F., & Cohen, D. I. Relation of the A-B distinction and trust-distrust sets to addict patients' self-disclosures in brief interviews. *Journal of Consulting and Clinical Psychology*, 1970, 34, 289-296.
- Berzins, J. I., Ross, W. F., & Friedman, W. H. A-B therapist distinction, patient diagnosis, and outcome of brief psychotherapy in a college clinic. *Journal of Consulting and Clinical Psychology*, 1972, 38, 231-237.
- Berzins, J. I., & Seidman, E. Subjective reactions of A and B quasi-therapists to schizoid and neurotic communications: A replication and extension. *Journal of Consulting and Clinical Psychology*, 1968, 32, 342-347.
- Berzins, J. I., & Seidman, E. Differential therapeutic responding of A and B quasi-therapists to schizoid and neurotic communications. *Journal of Consulting and Clinical Psychology*, 1969, 33, 279-286.
- Chartier, G. M. The A-B therapist variable: Real or imagined? *Psychological Bulletin*, 1971, 75, 22-33.
- Green, M. H. The effectiveness of A and B college males as interviewers, as response-biasers and as verbal conditioners with schizophrenic and neurotic patients (Doctoral dissertation, University of Kentucky, 1971). *Dissertation Abstracts International*, 1971, 33, 915B-916B. (University Microfilms No. 72-21, 462)
- Jackson, D. M. *Personality Research Form manual*. Goshen, N.Y.: Research Psychologists Press, 1967.
- King, D. G., & Blaney, P. H. Effectiveness of A and B therapists with schizophrenics and neurotics: A referral study. *Journal of Consulting and Clinical Psychology*, 1977, 45, 407-411.
- Matthews, J. G., & Burkhart, B. R. A-B therapist status, patient diagnosis, and psychotherapy outcome in a psychiatric outpatient population. *Journal of Consulting and Clinical Psychology*, 1977, 45, 475-482.
- McNair, D. M., Callahan, D. M., & Lorr, M. Therapist "type" and patient response to psychotherapy. *Journal of Consulting Psychology*, 1962, 26, 425-429.
- Mehrabian, A. Relationship of attitude to seated posture, orientation, and distance. *Journal of Personality and Social Psychology*, 1968, 10, 26-30.
- Phillips, L., & Rabinovitch, M. Social role and patterns of symptomatic behavior. *Journal of Abnormal and Social Psychology*, 1958, 57, 181-186.
- Welch, R. D. The effectiveness of A and B college males and models and interviewers with schizophrenic and neurotic patients (Doctoral dissertation, University of Kentucky, 1971). *Dissertation Abstracts International*, 1971, 32, 2413B. (University Microfilms No. 71-25, 923)
- Whitehorn, J. C., & Betz, B. J. A study of psychotherapeutic relationships between physicians and schizophrenic patients. *American Journal of Psychiatry*, 1954, 111, 321-331.

Received November 9, 1977 ■

Drug Abuse Patterns, Personality Characteristics, and Relationships With Sex, Race, and Sensation Seeking

Patricia B. Sutker, Robert P. Archer, and Albert N. Allain

Department of Psychiatry and Behavioral Sciences
Medical University of South Carolina

Interrelationships among sex, race, drug use patterns, and personality variables were examined in a sample of 84 chronic users of illicit drugs. Subjects were administered the Minnesota Multiphasic Personality Inventory, the Sensation Seeking Scale, and the Shipley Institute of Living Scale and were interviewed using the Background Information Questionnaire. Comparisons were made between sex and ethnic subgroups on personality and drug use variables using analysis of covariance and chi-square procedures for subjects classified into high-, medium-, and low-sensation-seeking groups. Blacks were characterized by lower levels of sensation seeking, less psychopathology, use of fewer drug categories, and later drug use than whites. Use and personality patterns among women differed little from those of men. Levels of sensation seeking were related to specific personality constellations, number of drug categories used, and motive for first alcohol use.

Research on the psychological characteristics of drug abusers has developed from attempts to describe and differentiate addicts from representatives of other clinically deviant categories. More recently, investigators have compared drug abuse subgroups (defined by race and sex) on personality dimensions or drug use patterns. Female and white drug abusers have been shown to demonstrate greater psychopathology than males and nonwhites (DeLeon, 1974; Olson, 1964), and ethnicity was found to be related to choice of drug type and variety used, particularly among men (Kaestner, Rosen, & Appel, 1977). Suffet and Brotman (1976) reported

lower rates of illicit drug use among women, but sex differences in drug use patterns have not been adequately specified.

Studies have also attempted to describe the association between personality characteristics and chronic drug use often without sex or race comparisons. Relationships have been demonstrated between sensation seeking and drug use patterns in college students (Zuckerman, Bone, Neary, Mangelsdorff, & Brustman, 1972) and hospitalized male veterans (Kilpatrick, Sutker, Roitzsch, & Miller, 1976). Combined elevations on Minnesota Multiphasic Personality Inventory (MMPI) Scales *Pd* and *Ma*, suggestive of exaggerated tendencies toward social nonconformity, have also been associated with chronic illicit drug use in men (Sutker & Allain, 1973; Zuckerman, Sola, Masterson, & Angelone, 1975) and women (Sutker & Moan, 1972), and relationships between drug choice and scores on sensation-seeking and MMPI dimensions have been described (Carrol & Zuckerman, 1977). For the most part, however, sex or race comparisons have been made independently, and their potential interactions in influencing drug use patterns or associated per-

This study was funded in part by General Research Support Grant No. 22620-A800 from the Medical University of South Carolina.

Special appreciation is expressed to Maggie Pike, Director of Odyssey House Louisiana and to Barry Pike, Director of the New Orleans Odyssey facility. The authors are also grateful to Walter Klimek and Caren Langford for assistance in data reduction and analysis.

Requests for reprints should be sent to Patricia B. Sutker, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina, 171 Ashley Avenue, Charleston, South Carolina 29403.

sonality characteristics have not been fully explored. The present study was designed to address this area of limited investigation and to examine relationships between levels of sensation seeking, drug use patterns, and personality characteristics among chronic users of illicit drugs.

Method

Subjects were 84 drug abusers in residential treatment at Odyssey House Louisiana and included 38 white men, 18 white women, 22 black men, and 6 black women, a breakdown representative of program composition. Roughly 57% were addicted to opiates at program entry, and the remaining subjects were continuous users of stimulants (19%), depressants (14%), psychedelics (5%), or other drugs (5%). Treatment admission was nonvoluntary in over 90% of the cases. Subjects were selected from residents referred for psychological assessment at program promotion to Level 2, one of six hierarchical treatment stages. This selection procedure was used to minimize variation in length of time in current treatment (average of 3 months). Criteria for subject inclusion in data analyses were signed informed consent, demonstrated ability to read test items, treatment residence for 2 months, and history of drug abuse exceeding 2 years. Means for combined race and sex groups on age, education, and Shipley Institute of Living Scale scores were 24.29 years, 11.30 years, and 115.55, respectively. Preliminary Race \times Sex analyses of variance showed no differences between subgroups in age, education, length of current treatment, or length of continuous drug use. Differences were found on the Shipley, with blacks producing lower scores than whites, $F(1, 80) = 26.68, p < .01$.

Instruments used for data collection were (a) the Sensation Seeking Scale (SSS), a forced-choice questionnaire that measures individual differences in preferred optimal level of stimulation and yields five subscale and total scores; (b) the MMPI, scored for the 3 validity, 10 clinical, and Special Scales *A*, *R*, and *Es* (*K*-corrected *T* scores); (c) the Shipley Institute of Living Scale, a measure of verbal comprehension and problem-solving skills; and (d) the Background Information Questionnaire (BIQ), a structured interview developed by us to acquire information about personal history and patterns of drug use (e.g., age at first drug use; number of drug categories ever used; reason for first drug, alcohol, or opiate use; first drug used; and drug of choice). First responses to reason for drug, alcohol, and opiate use were each classified in one of three categories defined by Naditch (1975): reluctant use from social pressure, use for therapeutic intent, and use for pleasure or curiosity.

Sex and race subgroups were compared on 7 quantifiable BIQ measures using analysis of variance and on 6 SSS and 16 MMPI variables using analysis

of covariance procedures with the Shipley score as the covariate; *F* tests for simple effects were performed where significant interactions were identified. Chi-square analyses were performed to assess relationships between sex and race and five categories of BIQ responses, including reason for first drug, alcohol, and/or opiate use; first drug used; and drug of preference. Total SSS scores were used to divide subjects into three SSS groups: (a) low ($n = 14$), with scores 1 *SD* below the mean SSS score of 43; (b) medium ($n = 57$), with scores between ± 1 *SD* of the mean; and (c) high ($n = 13$), with scores 1 *SD* or more above the SSS mean. Preliminary analyses of variance indicated Shipley score differences between SSS groups, $F(2, 81) = 3.70, p < .05$, and a greater frequency of blacks (64%) in the low SSS group, $\chi^2(2) = 9.94, p < .01$. There were no significant differences in education or sex distribution. Thus, race and Shipley scores were used as covariates in analysis of covariance comparisons of SSS groups on MMPI variables, whereas analyses of variance were performed to compare groups on BIQ dimensions. Fisher's least significant difference tests were used to evaluate significant between-groups differences. Chi-square analyses were performed to test relationships between SSS groups and drug use patterns.

Results

Drug abuse subgroups defined by race differed significantly on SSS variables, and whites scored higher than blacks on Thrill and Adventure Seeking, $F(1, 79) = 18.79, p < .01$, General Sensation Seeking, $F(1, 79) = 9.18, p < .01$, and Total SSS, $F(1, 79) = 5.53, p < .05$. Sex differences in sensation seeking were limited, with men scoring higher than women only on Thrill and Adventure Seeking, $F(1, 79) = 5.78, p < .05$. Race and sex MMPI comparisons showed no differences between men and women, but whites produced higher scores on Scales *F*, $F(1, 79) = 4.03, p < .05$, *D*, $F(1, 79) = 4.61, p < .05$, *Pa*, $F(1, 79) = 4.32, p < .05$, and *Pt*, $F(1, 79) = 5.04, p < .05$, than blacks.

White drug abusers were younger than blacks at time of first drug, $F(1, 80) = 11.18, p < .01$, and first opiate use, $F(1, 61) = 6.51, p < .05$, and had used drugs from a greater variety of categories, $F(1, 80) = 6.06, p < .05$. Blacks and whites differed in drug of preference, $\chi^2(1) = 14.36, p < .01$. Opiates and depressants were endorsed by 100% of blacks and 62% of whites, whereas stimulants and hallucinogens were preferred

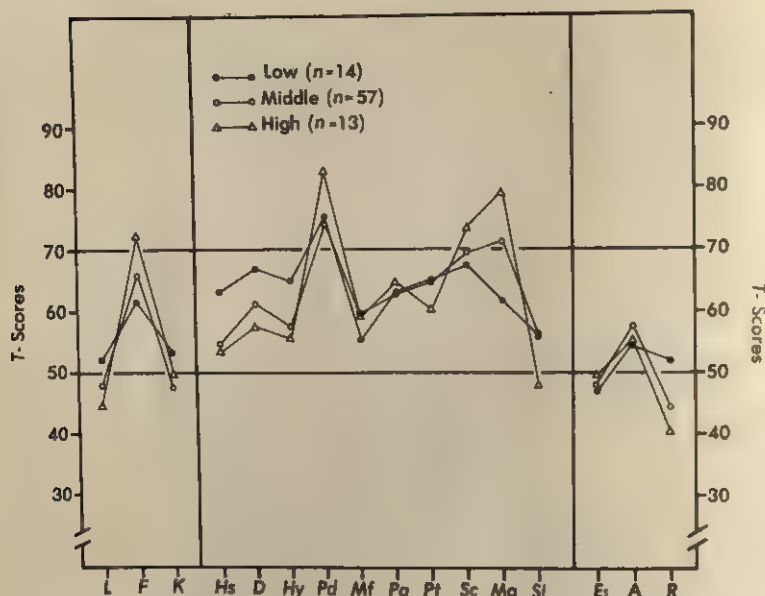


Figure 1. Mean Minnesota Multiphasic Personality Inventory profile patterns for low-, medium-, and high-sensation-seeking groups.

by 38% of whites and 0% of blacks. In male/female comparisons, men reported use of more drug categories, $F(1, 80) = 4.82$, $p < .05$, but women and men did not differ in reported drug preference. Race or sex differences were not significantly associated with reason for first drug, alcohol, or opiate use.

High, medium, and low SSS groups were characterized by variations in mean MMPI profile configurations (see Figure 1). Differences were found between groups on Scales *L*, $F(2, 79) = 4.22$, $p < .05$; *F*, $F(2, 79) = 3.19$, $p < .05$; *Hs*, $F(2, 79) = 3.49$, $p < .05$; *Hy*, $F(2, 79) = 4.50$, $p < .05$; *Pd*, $F(2, 79) = 4.23$, $p < .05$; *Ma*, $F(2, 79) = 10.00$, $p < .01$; *Si*, $F(2, 79) = 3.90$, $p < .05$; and *R*, $F(2, 79) = 8.93$, $p < .01$. Low sensation seekers scored higher than middle sensation seekers on Scales *L* and *Hs* ($p < .05$), *Hy* and *R* ($p < .01$) and high sensation seekers on *Hs* ($p < .05$) and *L*, *Hy*, and *R* ($p < .01$). High sensation seekers produced lower scores on Scale *Si* than those in low ($p < .05$) and middle ($p < .01$) groups and higher scores on Scales *F* and *Pd* than other groups ($p < .05$). High sensation seekers also produced more elevated scores on Scale *Ma* than medium sensation seekers ($p < .01$), who in turn produced

higher scores on *Ma* than those in the low SSS group ($p < .01$).

Sensation-seeking levels were significantly related to drug use patterns. High and middle sensation seekers reported earlier, $F(2, 81) = 3.15$, $p < .05$, and more varied, $F(2, 81) = 9.10$, $p < .01$, use of drugs than low sensation seekers. Although reason for first drug or opiate use and drug of choice did not vary as a function of SSS classification, reason for first alcohol use differed across groups, $\chi^2(4) = 12.96$, $p < .05$. Among low sensation seekers, 62% remembered their first use of alcohol as motivated by the influence of others, whereas 67% of high sensation seekers attributed initial use of alcohol to pleasure and curiosity.

Discussion

Findings indicate that race is an important factor to consider in understanding drug abuse phenomena, but gender may be of limited value in prediction of personality or drug use patterns for illicit drug users. Consistent with Kaestner et al. (1977), blacks demonstrated lower levels of sensation seeking and less psychopathology, reported use of fewer drug categories, showed preference for

depressants such as opiates over stimulants, and engaged in drug use later than whites. Less elevated scores on MMPI measures occurred among blacks despite the possibility that current MMPI norms, derived from white reference groups, exaggerate the *T*-score estimates of psychopathology for blacks (Gynther, 1972). In contrast to earlier research (DeLeon, 1974; Olson, 1964), results suggest that female drug abusers, in reference to their normative sex group, are no more psychologically deviant than men. Thus, the issue of sex-specific personality differences cannot be resolved without further comparisons across treatment and nontreatment conditions. Women and men also showed few dissimilarities on sensation-seeking measures with the exception of Thrill and Adventure Seeking, a cluster of items reflecting desire to engage in outdoor sports or activities involving speed or danger. Although women reported use of drugs from a fewer number of categories, no sex differences were found in age at first drug use, frequency of drug use, or drug preference.

Results support the hypothesis that there is a close relationship between sensation seeking, other personality dimensions such as sociopathy and neurotic involvement, and drug use patterns. High sensation seeking was related to use of more drug categories, earlier age at first drug use, and curiosity as a motive for initial alcohol use. Drug abusers classified as high sensation seekers scored higher on scales reflecting sociopathy, attitudinal deviance, and heightened activity and lower on measures indicating denial, hypochondriacal preoccupation, hysteria, and social introversion. Such individuals, relatively uninhibited by neurotic defenses, seem to be strongly motivated to increase external stimulation. In contrast, low sensation seekers produced higher elevations on measures of neurotic involvement, repression, and denial. Similar relationships among SSS and MMPI variables have been reported in correlational studies with prisoners (Blackburn, 1969) and alcoholics (Kish & Busse, 1969).

Present findings suggest that motives for drug use vary depending on such critical variables as race, sensation seeking, neurotic involvement, and sociopathy. It might be

hypothesized that chronic drug use is associated with exaggerated needs to attenuate unpleasant internal states or, conversely, to seek out external sources of stimulation. These assumptions provide a basis on which to match specific therapeutic packages to client personality characteristics and drug use patterns as well as a reasonable framework for investigating treatment outcome. For example, treatment of low sensation seekers might incorporate relaxation and social skills training to provide alternatives to drug use for reducing unpleasant internal states; high sensation seekers could be encouraged to identify activities and goals that provide gratifying and stimulating alternatives to the pharmacologic effects and concomitant life-styles of illicit drug use. Finally, motives for drug use and their relationships with such variables as race, sensation seeking, neurotic defenses, social introversion, and sociopathy should be explored systematically among drug experimenter, chronic user, treatment, and post-treatment populations.

References

- Blackburn, R. Sensation seeking, impulsivity, and psychopathic personality. *Journal of Consulting and Clinical Psychology*, 1969, 33, 571-574.
- Carrol, E. N., & Zuckerman, M. Psychopathology and sensation seeking in "downers," "speeders," and "trippers": A study of the relationship between personality and drug choice. *International Journal of the Addictions*, 1977, 12, 591-601.
- DeLeon, G. Phoenix House: Psychopathological signs among male and female drug-free residents. *Addictive Diseases: An International Journal*, 1974, 2, 135-151.
- Gynther, M. D. White norms and black MMPIs: A prescription for discrimination? *Psychological Bulletin*, 1972, 78, 386-402.
- Kaestner, E., Rosen, L., & Appel, P. Patterns of drug abuse: Relationships with ethnicity, sensation seeking, and anxiety. *Journal of Consulting and Clinical Psychology*, 1977, 45, 462-468.
- Kilpatrick, D. G., Sutker, P. B., Roitzsch, J. C., & Miller, W. C. Personality correlates of polydrug use. *Psychological Reports*, 1976, 38, 311-317.
- Kish, G. B., & Busse, W. MMPI correlates of sensation seeking in male alcoholics: A test of Quay's hypothesis applied to alcoholism. *Journal of Clinical Psychology*, 1969, 25, 60-62.
- Naditch, M. P. Relation of motives for drug use and psychopathology in the development of acute adverse reactions to psychoactive drugs. *Journal of Abnormal Psychology*, 1975, 84, 374-385.

- Olson, R. W. MMPI sex differences in narcotic addicts. *Journal of General Psychology*, 1964, 71, 257-266.
- Suffet, F., & Brotman, R. Female drug use: Some observations. *International Journal of the Addictions*, 1976, 11, 19-33.
- Sutker, P. B., & Allain, A. N. Incarcerated and street heroin addicts: A personality comparison. *Psychological Reports*, 1973, 32, 243-246.
- Sutker, P. B., & Moan, C. E. Personality characteristics of socially deviant women. Incarcerated heroin addicts, street addicts, and nonaddict prisoners. In J. M. Singh, L. H. Miller, & H. Lal (Eds.), *Drug addiction: Clinical and socio-legal aspects* (Vol. 2). New York: Futura, 1972.
- Zuckerman, M., Bone, R. W., Neary, R., Mangelsdorff, D., & Brustman, B. What is the sensation seeker? Personality trait and experience correlates of the Sensation Seeking Scales. *Journal of Consulting and Clinical Psychology*, 1972, 39, 308-321.
- Zuckerman, M., Sola, S., Masterson, J., & Angelone, J. MMPI patterns of drug abusers before and after treatment in therapeutic communities. *Journal of Consulting and Clinical Psychology*, 1975, 43, 286-296.

Received November 21, 1977 ■

Marital Satisfaction and Depression as Predictors of Physical Health Status

Robert L. Weiss and Barbara M. Aved
University of Oregon

This study investigated the multiple correlation between physical health status and a set of marriage-related "predictor" variables. Family practice physicians provided a sample of 104 married couples. Marital satisfaction, depression, number of visits to physician, and educational level were among the set of cross-validated "predictors" of reported physical health status. The correlation between physical health status and depression was significantly greater for wives than husbands. For wives, marital satisfaction and depression were related primarily through the uncontrolled variance in physical health status, whereas for husbands a significant relationship between marital satisfaction and depression remained for husbands when physical health status was partialled out. These findings support similar conclusions drawn by others.

Marital relationships provide a particularly attractive point of entry for the study of the relationship between behavior and health. Socially oriented conceptions of mental health focus either on interpersonal learning or on much broader psychosocial systems, such as collectives and communities. The marital unit provides the investigator with a vantage point within a minisystem from which attention can be directed downward toward individual or upward toward community determinants of behavior and adjustment. Nonetheless, empirical relationships between marital status and various conceptions of "health" have been studied only imperfectly (e.g., Crago, 1972; Vincent, 1973).

Some recently reported empirical relationships are as follows: (a) Marriage-related requests for professional mental health services run high; 58% of 2,000 consecutive

outpatient requests at a large university hospital were marriage related (Overall, Henry, & Woodward, 1974), whereas for psychological and social work settings, estimates run from 58% to 76% of all presenting problems. (b) Persons who have had multiple marriages, relative to the never married, show significantly less rated psychopathology (Overall, 1971). (c) Marital dissatisfaction may be part of a broader depressive spectrum disorder (Overall et al., 1974). The direction of causality between psychosocial variables and marital satisfaction has not been determined by these and the many other studies available in the literature. It is just as reasonable to assume that marital dissatisfaction causes depressive symptoms, poor job performance, or poor health as it is to assume that marital dissatisfaction results from these maladies.

The aims of this investigation were modest: We sought to establish the multiple covariations between reported physical health status and at least two face-valid person variables: marital satisfaction and depression. A similar approach, reported by Coleman and Miller (1975), determined that both self-reported and therapist-rated depression were significantly related to marital satisfaction, accounting for 14% and 23%

This article is based on a dissertation submitted in partial fulfillment of requirements for the PhD degree by the second author to the Department of Health Education, University of Oregon.

The authors wish to thank the physician members of the Lane County Academy of Family Practice for their cooperation in all phases of this study.

Requests for reprints should be sent to Robert L. Weiss, Department of Psychology, University of Oregon, Eugene, Oregon 97403.

of the variance in satisfaction ratings, respectively. Since physical health status and depression tend to covary (depression frequently involves somatic complaints), as do depression and marital satisfaction, the relationship between physical health status and marital satisfaction might be a spurious one.

The present study sought to establish the multiple correlation between a set of "predictor" variables and reported physical health status. Active patients of family practice physicians were selected for study, because physicians are often first-line providers of health services. They are an integral part of the psychosocial context of health status, since they frequently suspect that marital adjustment and depression, among other sources of distress, are concomitants (or even causal agents) of impaired health status. The multiple correlational analysis of health status and relationship variables was centered in this broader consumer-provider context. Physical health status has been a variable of psychological interest for those concerned with psychophysiological reactions and, more recently, life stress and somatic illness (cf. Holmes & Masuda, 1974; Rabkin & Struening, 1976; Rahe, 1974). Since the longitudinal study of changes in health status that occurs during marriage is impractical, the present study sought to provide cross-sectional data.

Method

Subjects

Family practice physicians at a regular meeting of the Lane County Academy of Family Practice were invited to participate in a study of the relationship between physical health status and variables related to marital satisfaction by providing names of couples who met the following criteria: (a) married for at least 2 years, (b) in a first marriage for both spouses, (c) between the ages of 22 and 48 years, and (d) wife not now pregnant. These criteria assured that couples were beyond the initial stages of marital accommodation and that medical problems associated with advancing age and/or pregnancy were excluded. The physicians were also asked to indicate whether, relative to their own practice, the persons they nominated were "overusers" or "normal users" of their services. The aim was to include representation of patients having psychosomatic involvements.

Four physicians participated by contributing a total of 280 persons (140 couples). (In almost all

instances, husbands and wives were patients of the same physician.) Each physician drafted a letter to his own patients, explaining that the study was "surveying adults about their health patterns," indicating an interest in identifying factors of family health potentially related to prevention, and assuring them that refusal to participate would not influence their relationship with their physician.

From the initial pool of 140 couples who had been contacted by letter, 135 were subsequently contacted by phone; 104 couples completed usable forms. Participation was refused in 23 cases; 5 couples had separated or divorced, 1 wife was pregnant, and 2 couples provided incomplete data. In the final sample, 18% of the subjects were designated by their own physicians as overusers (22% of the refusals had this designation). Individual physicians contributed from 17% to 38% of the final sample of 104 couples.

The typical subject was 30 years old, had had 2 years of college, and had been married for 8 years. Fifty-four percent of husbands and 24% of wives held white-collar and professional positions; 4% of husbands and 61% of wives were unemployed; 66% of the sample had two or more children. Couples were predominantly Protestant (57%), with 19% Catholics, 6% other, and 18% none. The resulting sample was a well-educated, possibly upper-middle class, group of couples who were rearing preadolescent children.

Measures

Physical health status. The Cornell Medical Index (CMI) was used as the measure of physical health status (Brodman, Erdman, Lorge, & Wolff, 1952); it consists of 195 items typically covered in an intensive medical history interview, ranging from systemic (sensory, respiratory, urogenital, etc.) to vague complaints associated with psychosomatic disorders. The questions sample bodily symptoms, past illnesses, family history, and affective states. The score is the total number of affirmative responses to the questions. In a sample of women scheduled for cancer surgery, both their preoperative and postoperative CMI scores were related to staff ratings of postsurgery invalidism ($r = .49$ and $.52$, respectively, $p < .05$); the pre-surgery to postsurgery CMI scores were also related ($\rho = .67$, $p < .01$; Bard & Waxenburg, 1957).

Self-Rating Depression Scale. The Zung (1965) Self-Rating Depression Scale (SDS) was used to assess depression. It consists of 20 self-statements, half of which are worded affirmatively and half, negatively. The SDS has been used widely in depression research for detecting the so-called "hidden depressions." Individual scores can range from 20 to 80 on this scale. Zung (1974) reviewed the literature on the reliability and concurrent and predictive validity of this measure. A number of studies have shown that for age-matched depressive patients and normals, 88% of the patients were detected by the SDS, whereas 12% of the normals were detected as "false positives."

Marital satisfaction. The Locke-Wallace Marital Adjustment Test was used to assess marital satisfaction and has been used widely in the literature as a reliable measure of marital satisfaction (Coleman & Miller, 1975; Locke & Wallace, 1959; Weiss & Margolin, 1977). Each spouse answers individually; scores of 100 or more generally indicate marital satisfaction, and scores below 100 indicate marital distress.

Activity level inventory. A 28-item version of the Inventory of Rewarding Activities, developed by the Marital Studies Program, University of Oregon, was used to assess how couples spend their leisure time. The scale reflects activities in home and community equally (e.g., "watching TV for more than 1 hour" and "go to a dance or party"). Respondents indicate whether during the last week each activity was engaged in alone or with spouse. Four activity scores were obtained: Home with spouse, Home without spouse, community with spouse, and community without spouse.

Questionnaire. In addition to the six measures described above, information on five demographic variables (age, number of annual visits to physician, years married, years residence in county, and education) was obtained.

Procedure

Persons to whom individual physicians had sent letters introducing the study were contacted by telephone and were given additional information. If they agreed to participate, an in-home meeting was then scheduled.

All data were collected in this meeting, which lasted approximately 45 minutes. The same interviewer (the second author) a registered nurse, conducted all interviews with respondents within a 3-month period.

When all interviews had been completed, the master lists of patients disclosing the initial "user" status designation were made available to the interviewer; the authors did not inform physicians which of their patients had or had not participated.

Results

Husband-Wife Comparisons

Wives, compared to husbands, on the average made significantly ($p < .01$ for all comparisons) more visits (annually) to physician, had more medical complaints (higher CMI scores), and reported engaging in more home alone activities. Husbands and wives did not differ significantly in mean Locke-Wallace marital satisfaction scores ($M_s = 111.4$ and 113.1 , respectively). The correlation between spouses' marital satisfaction scores was $.65$ ($p < .001$).

As an internal validity check, the 19 couples identified by their own physicians as "overusers" of medical services were compared to the 85 designated as "normal user" couples. In all comparisons, the females in the overuser group were significantly different from their counterparts in the normal group, whereas the males in both groups were comparable. As defined by their individual physicians, overuser relative to normal user females, on the average, scored 10 points higher on the CMI, 4 points higher on the SDS, and 7 points lower on marital satisfaction, and they were reported as having made twice as many visits to their physicians.

These comparisons between males and females are generally consistent with previous findings: Females reported more physical health concerns, made more visits to physician, and tended to score as somewhat more depressed on the SDS.

Predictors of Physical Health Status (CMI)

The 11 predictor variables were combined through multiple linear regression equations, with CMI as the dependent variable. Although the combination of self-report and demographic variables was selected for presumed relevance to marital interaction, multiple regression is particularly sensitive to fluctuations among correlations within successive samples. Consequently, the sample of 104 couples was divided at random into equal subsets of $N_A = N_B = 52$ couples. The neutral designation, Sample A and B, is preferred, since the dichotomy was formed after all data had been collected.

The results of major interest are presented separately in Table 1 for sex within each A and B subsample, as well as for subsample totals for each sex (total $N = 104$).

The zero-order correlations between each of the 11 predictors and CMI are listed for husbands and wives separately for subsamples and for totals, thereby allowing comparisons of within- and between-subsample fluctuations.

The multiple regression coefficients that resulted in each instance for the 11 predictors are listed below the zero-order correlations. The R^2 values indicate the amount

Table 1
Zero-Order, Multiple Correlation, and Validity Coefficients for Regression of CMI on Predictor Variables

Predictor	Husbands			Wives		
	A	B	Total	A	B	Total
SDS	395	374	379 ^a	650	705	673 ^a
Marital satisfaction	-470	-464	-468	-569	-371	-459
AHS	-166	-148	-137	-335	-171	-258
AHS	-017	093	-034	-035	-208	-124
ACS	025	071	-022	033	-194	-071
ACS	-013	-047	-028	-045	-238	-139
Age	063	106	082	-225	-273	245
Years married	267	012	147	-165	-109	-143
No. visits	392	375	384	278	512	346
Years in county	146	023	083	392	018	223
Education	-279	-143	-216	-263	-534	-394
R^b	664	659	629	823	840	777
R^2	44	43	395	68	70	604
Validity coefficient r	484	487	(485)	668	568	(620)

Note. $N_A = N_B = 52$. Decimals have been omitted. Validity coefficients are all significant at the .001 level. Averaged values appear in parentheses. For 50 df , $r_{.95} = .273$; for 102 df , $r_{.95} = .195$. CMI = Cornell Medical Index; SDS = Self-Rating Depression Scale; AHS = activities at home with spouse; AHS = activities at home without spouse; ACS = activities in the community with spouse; ACS = activities in the community without spouse.

^a Difference between these zero-order correlations is significant at the .002 level.

^b For Subsample A = B $df = 11, 40$; Total(s) $df = 11, 92$. F values for Husbands A, B, and Total: $F = 3.23, 2.79, 5.47$. F values for Wives A, B, and Total: $F = 7.72, 8.72, 12.79$. For all F values ≥ 3.23 , $p < .01$; $F = 2.79$, $p < .05$.

of CMI variance accounted for by the set of predictors.

Each of the multiple correlations was significant; the range of CMI variance varied from a low of 40% to a high of 70%. From the zero-order correlations in Table 1, it can be seen that the predictive ability of SDS, the depression variable, was most different for husbands and wives. (The significance of the difference between these correlations, based on $\sigma_{r_1-r_2} = .141$, yielded $p = .002$.)

For the two husband subsamples, the "best" replicated predictors included marital satisfactions, visits, and SDS; those for the wives included SDS and marital satisfaction, similar to husbands. Nonreplicating predictors for the wives included education, visits, years in county, and activity in community without spouse.

Finally, the last row of Table 1 lists the validity coefficients determined by "double cross-validation"; the beta regression

weights, derived from each within-sex subsample, were applied reciprocally to the other. For example, the betas for Sample A wives were applied to Sample B wives, and those for Sample B were applied to Sample A. The Pearson product-moment correlation coefficients between the predicted CMI and observed CMI scores thus yield a validity coefficient based on the adequacy of one set of beta weights to predict from the raw data of a different sample. This particular choice of cross-validation underestimates the goodness of fit, since only one half of the data is used to determine the weights.

From Table 1 it can be seen that all validity coefficients were highly significant, although those for husband samples were substantially lower than the estimates for wives.

To further specify the relationship between CMI (1), marital satisfaction (2), and SDS (3), the partial correlation ($r_{12.3}$) was calculated for husbands and wives separately. The respective partial correlations were $-.35$

and $-.28$ ($ps < .01$). From Table 1, the comparable zero-order correlations are $-.468$ and $-.459$.) Reported physical health status and marital satisfaction were mediated, in part, by variations in depressiveness scores; the partial correlations remained statistically significant.

When the relationship between marital satisfaction and depressiveness was statistically controlled for differences in physical health status scores, the partial correlations dropped from their respective zero-order correlations of $-.486$ and $-.401$ ($ps < .01$) to $r_{12.8} = -.334$ ($p < .01$) and $-.159$ (ns), for husbands and wives, respectively. The CMI and SDS share many physical items in common (e.g., "I have trouble with constipation."). The significant residual correlation between SDS and marital satisfaction suggests that marital (dis)satisfaction and depressive affect are likely to be found together for husbands, although the magnitude of the correlation was small. For wives it appears that common variance from the physical symptom items accounts for the CMI \times MS relationship, which is understandable given the size of the zero-order correlation between CMI and SDS in Table 1.

Discussion

This has been an investigation of the network of covariations between physical health status on the one hand and marriage-related variables on the other. It differed from others (e.g., Coleman & Miller, 1975) by (a) drawing nonclinical spouses from the context of an ongoing consumer-provider relationship, namely, patients and their family practice physician providers, and (b) providing a cross-validation of the correlates of health status. Well-educated couples in first marriages who reported above average marital satisfaction served as respondents. Physicians identified a relatively small number of their patients as overusers of their services to ensure inclusion of typical "psychosomatic" cases. (Descriptively, these cases were women who differed significantly from their counterparts on all relevant variables, e.g., annual visits, CMI, and SDS.)

The "best predictors" to CMI among the zero-order correlations that replicated within same-sex subsamples were SDS, marital satisfaction, visits, and to a lesser extent education. The amount of physical health status variance accounted for by SDS variance differed significantly for husbands and wives (14% and 45%, respectively).

When all 11 predictors were combined into multiple regression equations, the amount of CMI variance accounted for increased substantially, ranging from 40% to 60% for husbands and wives, respectively. Overall, CMI scores were better predicted for wives than for husbands.

Among the best predictors of physical health status, depression and marital satisfaction are the most interesting. (Persons reporting greater numbers of physical symptoms tended to see their physicians more than others, and those with more years of education reported fewer physical symptoms.)

Whereas the relationship between health status and marital satisfaction was only slightly reduced when unplanned variations in depression were statistically controlled, the correlation between marital satisfaction and depression for wives became insignificant when health status variance was controlled.

Coleman and Miller (1975) reported similar results for an older, less-educated sample of mental health clinic outpatients and their spouses. (The same pattern of mean differences between husbands and wives on depression and satisfaction ratings was observed in the two investigations.) For combined samples of husbands and wives, the relationship between marital satisfaction and depression was $r = -.38$ ($p < .01$) and $r = -.43$ ($p < .01$) in the Coleman and Miller and the present studies, respectively. They also found a similar pattern of difference between marital satisfaction and depression relationships for husbands and wives (i.e., $rs = -.66$, $p < .01$, and $-.25$, $p = .10$, respectively). In the present results the Marital Satisfaction \times SDS relationship became insignificant for wives when the uncontrolled variations in physical health status were accounted for. (In both studies a significant inverse relationship was found between husbands' mari-

tal satisfaction and wives' depression scores, but in neither case were wives' depression and husbands' marital satisfaction scores significantly related.)

Coleman and Miller (1975) concluded that marital satisfaction and depression for females were essentially unrelated, and the present results lend some support to this view. Bernard (1972), on the other hand, concluded that wives suffer far greater mental health hazards and that they present a far worse clinical picture than do husbands. Drawing mainly on sociological studies (e.g., Knupfer, Clark, & Room, 1966), she stressed the double standard of mental health that essentially builds into "our standards of mental health for women the defects necessary for successful adjustment in marriage" (p. 52). More recently Overall et al. (1974) have presented statistical evidence that both a family history of marital discord and what might be labeled a "depressive spectrum disorder" independently predict the likelihood of outpatients' complaints of marital discord.

All of the available data are based on correlational analyses. This approach may be helpful in suggesting which aspects of relationship living have significance for physical health status. A direct test is needed whereby marital therapy can be shown to improve physical health status of the partners; depending on one's point of view, this may occur more for husbands than for wives. The present investigation is merely another step toward specifying the probable contribution of separate variables.

References

Bard, M., Waxenberg, S. Relationship of Cornell Medical Index responses to postsurgical invalidism. *Journal of Clinical Psychology*, 1957, 13, 151-153.
Bernard, J. *The future of marriage*. World: New York, 1972.
Brodman, K., Erdman, Jr., A. J., Lorge, L., & Wolff, H. C. The Cornell Medical Index-Health Questionnaire: VI. The relationship of patients' complaints to age, sex, race, and education. *Journal of the American Medical Association*, 1952, 147, 112-119.

naI of the American Medical Association, 1952, 147, 112-119.
Coleman, R. E., & Miller, A. G. The relationship between depression and marital maladjustment in a clinic population: A multitrait-multimethod study. *Journal of Consulting and Clinical Psychology*, 1975, 43, 647-651.
Crago, M. A. Psychopathology in married couples. *Psychological Bulletin*, 1972, 77, 114-128.
Holmes, T. H., & Masuda, M. Life change and illness susceptibility. In B. S. Dohrenwend & B. P. Dohrenwend (Eds.), *Stressful life events: Their nature and effects*. New York: Wiley, 1974.
Knupfer, G., Clark, W., & Room, R. The mental health of the unmarried. *American Journal of Psychiatry*, 1966, 122, 841-851.
Locke, H., & Wallace, K. Short marital-adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living*, 1959, 21, 251-255.
Overall, J. E. Associations between marital history and the nature of manifest psychopathology. *Journal of Abnormal Psychology*, 1971, 78, 213-221.
Overall, J. E., Henry, B. W., & Woodward, A. Dependence of marital problems on parental family history. *Journal of Abnormal Psychology*, 1974, 83, 446-450.
Rabkin, J. G., & Struening, E. L. Life events, stress, and illness. *Science*, 1976, 194, 1013-1020.
Rahe, R. H. Life change and subsequent illness reports. In E. Gunderson & R. Rahe (Eds.), *Life stress and illness*. Springfield, Ill.: Charles C. Thomas, 1974.
Vincent, C. *Sexual and marital health: The physician as a consultant*. New York: McGraw-Hill, 1973.
Weiss, R. L., Hops, H., & Patterson, G. R. A framework for conceptualizing marital conflict, a technology for altering it, and some data for evaluating it. In F. W. Clark & L. A. Hamerlynck (Eds.), *Critical issues in research and practice: Proceedings of the Fourth Banff International Conference on behavior modification*. Champaign, Ill.: Research Press, 1973.
Weiss, R. L., & Margolin, G. Marital conflict and accord. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook for behavioral assessment*. New York: Wiley, 1977.
Zung, W. A self-rating depression scale. *Archives of General Psychiatry*, 1965, 12, 63-70.
Zung, W. The measurement of affects: Depression and anxiety. *Psychological Measurements in Psychopharmacology*, 1974, 7, 170-188.

Received November 22, 1977

Sexual Fantasies of Females as a Function of Sex Guilt and Experimental Response Cues

Denise Moreault and Diane R. Follingstad
University of South Carolina

This study investigated the effects of response cues (erotic, romantic, or neutral) and level of sex guilt on the self-reported sexual fantasies of females. Undergraduates completed a sex guilt inventory, a mood adjective checklist, a fantasy theme checklist, and ratings of their affective responses and physiological arousal associated with the writing of the fantasies. High sex guilt females preferred fantasy themes indicating a lack of responsibility for engaging in sexual interaction. Subjects in the erotic fantasy condition wrote more explicit fantasies and described more varied content. Arousal seemed to be affected by the response cuing in the predicted direction but not by the subjects' guilt levels. Sex guilt level seemed to be a better predictor of affective responses, such as guilt and embarrassment, than the response cuing. The results suggest that sexual fantasy behavior may be part of a cluster of sexual behaviors governed by an individual's level of sex guilt. The demonstration that fantasy production seemed to be influenced by situational demands has implications for collection and use of fantasy information by both clinicians and researchers.

Until recently, sexual fantasy behavior in females has remained a relatively unexplored area except among clinicians studying cases of sexual deviation (e.g., Abraham, 1922; Eidelberger, 1945; Sterba, 1921; Wulff, 1942). It has been viewed as pathological or at least symptomatic of poor heterosexual relations, especially when the fantasies included other than conventional heterosexual practices (Hollender, 1970). Only one experimental study (Hariton & Singer, 1974) in this area established the normalcy of sexual fantasies, substantiating Kinsey, Pomeroy, Martin, and Gebhard's (1953) normative demographic data. This study, however, only correlated a few personality variables with fantasy, and predictability from their data was limited.

The limited information regarding females' sexual fantasies has been complicated by a treatment of sexual fantasy as a homogenous entity in which a sample of individuals are simply asked whether or not they engage in

fantasy during sexual experiences. No investigation to date has selected subjects along a personality trait dimension to determine whether certain individuals are more likely to fantasize or are affected differently by fantasizing. In addition, specific operationalized parameters of fantasy behavior (e.g., themes, vividness, number of acts and organs portrayed) that potentially could yield more precise information have been sacrificed for simple assessment of the presence or absence of fantasies. Researchers attempting to collect data regarding sexual fantasies have also neglected to study the circumstances or conditions under which fantasies are produced and disclosed to others, thus ignoring the demand characteristics of the situation. The purpose of this study in assessing females' sexual fantasy behavior was to determine whether fantasies were tied to a personality trait and/or were under situational stimulus control.

A personality dimension that has been shown to predict a variety of sexual behavior patterns is the construct of sex guilt as developed, defined, and measured by Mosher (1966). Sex guilt has been defined as "a

Requests for reprints should be sent to Diane Follingstad, Department of Psychology, University of South Carolina, Columbia, South Carolina 29208.

generalized expectancy for self-mediated punishment for violating or for anticipating violating standards of proper sexual conduct" (Mosher & Cross, 1971, p. 27). Since sexual fantasies are likely to be related to other sexual behaviors, the level of sex guilt in individuals was expected to influence the content and explicitness of fantasies. Subjects scoring high in sex guilt compared to subjects low in sex guilt have been found to produce or recall less information relating to sexuality (Galbraith & Mosher, 1968; Langston, 1973; Schwartz, 1973), to evaluate less conventional and explicit sexual activity negatively (Mosher, 1973; Ray & Walker, 1973), and to report more guilt and embarrassment following exposure to sexual stimuli (Mosher & Greenberg, 1969; Schill & Chapin, 1972). Based on the literature regarding the effects of levels of sex guilt, the fantasies of high sex guilt (HSG) subjects were expected to be fewer in number, shorter, and less vividly remembered; to exhibit less thematic variety; to be more restricted in terms of specific sexual acts and organs described; and to contain romantic heterosexual themes. Following the writing of their own fantasies, subjects with high levels of sex guilt were expected to report more guilt and embarrassment than low sex guilt (LSG) subjects.

There have been conflicting data regarding how level of sex guilt affects sexual arousal. Mosher (1973) suggested that most females experience arousal regardless of their sex guilt level, whereas Schill (1972) suggested that subjects with little sex guilt report more arousal. It was predicted that subjects would report arousal to sexual fantasy stimuli independent of their level of sex guilt. In light of Schmidt's (1975) findings that females report a greater degree of arousal when rating genital sensations than when rating general arousal, sexual arousal needed to be measured by a variety of ratings.

Some investigators have been cognizant of the fact that the atmosphere and context in which self-disclosure takes place has an effect on the data that subjects produce (Galbraith & Mosher, 1968; Mosher, 1965). The stimulus conditions under which sexual fantasies are reported have important implications for

research and clinical practice, which relies on self-reported fantasy of subjects. No study, however, has attempted to manipulate the conditions under which fantasies are produced to determine which conditions are more conducive for, and differentially predictive of, the quality and quantity of subjects' fantasy production. Routh, Warehime, Gresen, and Roger (1973) found that giving subjects explicit instructions to write sexy stories as opposed to simply writing stories produced sexier stories. This would imply that fantasy content is under the influence of instructional set and experimental demand characteristics. Also, no researchers have investigated the effect of a romantic stimulus devoid of explicit sexual content on the sexual responsiveness of subjects (Schmidt, 1975).

Based on the research regarding instructional set, the following hypotheses were tested in relation to sexual fantasy production: (a) Erotic (i.e., sexually explicit) examples of fantasies would result in more and longer fantasies, more explicit imagery, and more variation in thematic content than romantic examples; (b) sexual fantasies would be longer, more varied, and more explicit for those females given sexual fantasy examples than those given a nonsexual fantasy stimulus; (c) individuals in a neutral fantasy condition would report less guilt and embarrassment than those in the erotic and romantic stimulus conditions, although erotic fantasy cues would produce a higher degree of sexual arousal in females than either romantic or neutral cues; and (d) a purely romantic stimulus would be less threatening than an erotic sexual stimulus for females with high sex guilt, resulting in more fantasy production with a greater variety of themes but less sexually explicit content.

The current literature on sexual fantasizing in clinical settings has been concerned with using fantasy to change deviant sexual behaviors (e.g., Evans, 1968; Marquis, 1970; Marshall, 1973) and assessing its effect in increasing sexual pleasure and treating sexual dysfunctions (e.g., Kaplan, 1974; Kline-Graber & Graber, 1975). Knowing the effects of level of sex guilt on a female's fantasizing behavior as well as the conditions most conducive for producing sexual fantasies should

contribute to the effective application of fantasy techniques in clinical settings.

Method

Subjects

Undergraduate females, ranging in age from 17 to 25, volunteered from a variety of psychology classes in a large southern university. One hundred fifty-three subjects were randomly assigned initially to three experimental conditions, receiving either erotic, romantic, or nonsexual fantasy examples. The final sample of 90 heterosexual subjects consisted of the 30 females in each condition who had the highest and lowest levels of sex guilt as measured by an inventory tapping self-reported guilt feelings. Ethical guidelines with regard to use of human subjects were followed.

Experimenter

The experimenter delivered the instructions and absented herself while the subjects completed the measures. To decrease the fear of external censure, the experimenter attempted to present herself as a nonevaluative and scientific individual (Galbraith & Mosher, 1968; Mussen & Scodel, 1955).

Measures

Mosher Forced-Choice Guilt Scale—Female Form. The Sex Guilt subscale consists of 20 Likert-type items with four response options. Split-half reliability of the scale produced a correlation of .95. Convergent and discriminative validity of the Mosher guilt scales have been demonstrated (Mosher, 1966; Mosher & Cross, 1971) with the Sex Guilt subscale being significantly different from the other guilt measures.

Fantasy essays. Subjects were asked to write down all the personal sexual fantasies that they could recall having experienced.

Fantasy Theme Checklist. This checklist was derived from the list developed by Hariton and Singer (1974) and was expanded to include themes reported in a compilation of women's sexual fantasies by Friday (1973). Subjects checked any of the 22 fantasy themes that they had experienced as part of their own fantasies.

Mood Adjective Check List (MACL). This measure of 49 adjectives was originally developed by Nowlis (1965) to prompt subjects to label their changing moods and was modified by Mosher and Greenberg (1969) to include 7 adjectives measuring sexual guilt and 7 adjectives measuring sexual arousal. Previous studies have demonstrated that changes in guilt and arousal as measured by adjective ratings are a viable measure of change due to experimental manipulation (Mosher & Greenberg, 1969; Okel & Mosher, 1968).

Sexual arousal, vividness and embarrassment rat-

ings. Subjects rated whether or not they experienced arousal, their subjective degree of sexual arousal on a 7-point scale, and the strength of genital and/or breast sensations in response to the writing of their fantasies. Embarrassment as a result of the fantasy reporting and the vividness (i.e., ability to remember facts and details clearly) of the fantasies while recording them were rated on a 5-point scale.

Experimental Conditions

Subjects in the three stimulus conditions were presented either with three erotic, three romantic, or three neutral examples of fantasies prior to writing their own fantasies. Heterosexual fantasies for the erotic (i.e., sexually explicit) condition were adapted from Friday's (1973) book, which reports fantasies solicited from the general population. Romantic fantasies, focusing on an emotional heterosexual relationship without mentioning sexual acts or organs, were solicited from undergraduate females. Nonsexual fantasies were solicited from other undergraduate females who wrote fantasies containing no sexual content, such as career aspirations or adventures.

Several fantasies of each type were given to a panel of five female judges (ranging in age from 23 to 31) prior to the experimental sessions and were rated by the judges along the dimensions of erotic, romantic, or personally pleasing for the three fantasy conditions mentioned above. The three fantasies rated highest in each condition were used as the experimental stimuli.

Procedure

Subjects in each experimental group were given the measures in three phases during a 1½-hour experimental session. Approximately 25 subjects were present at each session. In the first phase, subjects were administered the Mosher Forced-Choice Guilt Scale and the MACL. Following this, subjects were informed that the study involved reading and writing sexual materials. Subjects were informed that they could leave at any time.

To promote a permissive atmosphere, the second phase began with instructions informing subjects that sexual fantasies and thoughts were normal, usual experiences. The experimental stimuli were introduced with an explanation that reading examples of fantasies might put subjects in the mood for writing their own. Subjects were given the set of examples appropriate to the condition that they were in, and they returned the stimulus materials to the experimenter before proceeding. Explicit instructions of what constituted sexual fantasies were given in order that subjects would understand that either detailed or vague mental scenes, pictures, or stories that they had experienced would be appropriate to report. They were also told that these fantasies were to be written down in as much detail as possible.

Table 1
Means and Standard Deviations for Each Condition by Sex Guilt Level

Dependent variable	Condition					
	Erotic			Romantic		
	HSG			HSG		
	LSG	M	SD	LSG	M	SD
Length of fantasy	271	127	226	165	125	161
	428	183	298	338	131	261
Total words	2.86	2.13	2.20	4.73	4.48	2.60
No. fantasies	1.80	1.37	.86	.26	.80	.33
No. sex organs	4.20	1.21	2.86	1.66	1.05	1.40
No. sex acts	2.13	.99	1.20	1.53	.83	.86
Variety of content	9.21	4.44	6.40	8.40	3.35	7.26
Themes checked	1.27	.46	2.73	1.60	.74	2.20
Embarrassment	4.07	.70	2.87	3.40	.91	3.13
Vividness	4.40	1.96	4.00	3.47	1.73	3.20
Sexual arousal	.93	.26	.73	.67	.49	.47
Presence of sensations	3.13	1.06	2.40	2.13	1.19	2.13
Genital	2.07	.96	1.67	1.28	.41	1.33
Breast	12.1	3.54	13.8	11.6	4.41	12.7
Preguilt	10.6	4.11	12.0	10.8	3.95	12.4
Postguilt	12.8	4.47	12.5	12.2	4.88	11.5
Prearousal	21.3	3.86	16.3	17.6	4.12	15.7
Postarousal						
				Neutral		
				LSG	M	SD
				200	127	196
				414	128	339
				4.06	2.40	3.13
				.53	.99	.40
				2.00	1.31	2.06
				1.46	.64	1.33
				7.40	2.50	5.86
				1.53	.91	2.13
				3.26	1.03	2.60
				2.60	1.64	2.13
				.34	.49	.20
				1.53	.91	1.27
				1.06	.26	1.00
				11.1	3.81	15.3
				10.9	4.41	12.2
				11.6	5.86	11.5
				14.5	4.62	13.4
						3.15

Note. *n* = 15 for all groups, LSG = low sex guilt; HSG = high sex guilt.

Table 2

Summary Table of F Values by Fantasy Example Condition and Level of Sex Guilt

Dependent variable	Guilt (A)	Condition (B)	A × B
Length of fantasy	.47	4.01*	.29
Total words	9.75**	2.48	.36
No. fantasies	4.91*	1.71	.64
No. sex organs	2.56	9.48***	2.15
No. sex acts	3.85*	21.32***	2.63
Variety of content	3.57*	3.57*	2.70
No. themes checked	6.39**	1.19	.48
Embarrassment	23.57***	.28	2.49
Vividness	13.94***	2.66	2.01
Degree of arousal	1.15	9.09***	.02
Presence of sensations	3.58	12.15***	.05
Genital	2.24	12.61***	.92
Breast	.59	11.96***	1.17
Guilt* (A)	5.12*	.10	.28
Arousal* (B)	2.64	3.40*	.40
A × Time	1.13	.86	1.62
B × Time	3.86*	3.53*	1.08

* Measured by the Multiple Adjective Check List.

* $p < .05$.** $p < .01$.*** $p < .001$.

During the third phase, the following measures were administered: the Fantasy Checklist; the sexual arousal, vividness, and embarrassment ratings; and a second administration of the MACL. Subjects were debriefed after the experiment concluded.

Scoring of Subjects' Fantasies

Subjects' fantasies were scored according to (a) number of sexual fantasies; (b) length of longest sexual fantasy; (c) total number of words written; (d) explicitness of the sexual fantasy (i.e., the total number out of 7 possible sex organs and 12 possible sex acts that were mentioned either in physiological, medical, or slang terminology in the fantasy containing the highest frequency of acts and organs mentioned); and (e) total of 4 possible content areas of sexual fantasies (i.e., typical heterosexual activities; heterosexual oral-genital contact; group sex; rare or unusual sexual acts).

Results

Eighteen subjects were dropped from the original sample; 8 subjects reported having engaged in homosexual activities, 6 test records were incomplete, 1 subject was over the required age limit, and 3 subjects left the experiment after the nature of the study was revealed. The 15 subjects scoring highest and the 15 scoring lowest on sex guilt in each of

three experimental conditions comprised the final sample ($N = 90$). Ninety-eight percent of the sample wrote at least one sexual fantasy, and all subjects checked at least two fantasy themes.

In this 2×3 factorial design, the means of the HSG and LSG subjects across fantasy stimulus conditions were as follows: LSG in the erotic condition = -52.6; LSG in the romantic condition = -58.5; LSG in the neutral condition = -53.2; HSG in the erotic condition = -1.2; HSG in the romantic condition = -1.5; and HSG in the neutral condition = -.4. One-way analyses of variance (ANOVAS) performed on the means of sex guilt scores for LSG and HSG subjects across experimental conditions revealed no significant differences, supporting the random assignment of subjects to conditions and lending confidence to the assumption that any differences in dependent variables can be viewed as a result of the experimental manipulation and subjects' sex guilt level.

The hypothesis that HSG and LSG subjects would differ in the characteristics of their sexual fantasies was largely supported (see Tables 1 and 2 for means, standard deviations, and F values). One-way ANOVAS

Table 3

Chi-Square Results of Differences Between Sex Guilt Groups in Descriptions of Sex Organs and Sex Acts in Their Sexual Fantasies

Sex organ	χ^2	Sex act	χ^2
Breast	.05	Kissing	1.73
Penis	6.69**	Manipulation of female breast	2.19
Clitoris	5.94**	Manipulation of female genitals	.14
Vagina	.54	Intercourse	
		Ventral-ventral	4.84*
		Oral contact with male genitals	3.04
		Oral contact with female genitals	3.42
		Homosexuality	3.60

Note. For all significant differences, low sex guilt subjects were more explicit than high sex guilt subjects. Cell frequencies for the remaining three sex organs and five sex acts were too small to test. $df = 1$ in all cases.

* $p < .05$.

** $p < .01$.

indicated that LSG subjects demonstrated less sexual inhibition than HSG subjects in terms of sexual explicitness, expressiveness, and responsiveness to the stimuli: for total words, $F(2, 84) = 9.75$, $p < .01$; for total number of fantasies, $F(2, 84) = 4.91$, $p < .05$; for number of sex acts, $F(2, 84) = 3.85$, $p < .05$; for variety of content, $F(2, 84) = 3.57$, $p < .05$; and for number of fantasy themes checked, $F(2, 84) = 6.39$, $p < .01$. As predicted, there were no significant differences between HSG and LSG subjects on level of arousal or breast and genital sensations, even though LSG subjects increased the amount of arousal experienced significantly more from pre-MACL to post-MACL than HSG subjects, $F(2, 84) = 3.86$, $p < .05$. More embarrassment was reported by HSG subjects, $F(2, 84) = 23.57$, $p < .001$, and they also reported experiencing their fantasies less vividly, $F(2, 84) = 13.94$, $p < .001$ (see Table 2). LSG subjects reported experiencing less guilt following the sexual stimulation (postguilt) than HSG subjects, $F(2, 84) = 5.12$, $p < .05$ (see Tables 1 and 2), although HSG subjects in comparison to

the LSG subjects did not significantly increase the amount of guilt (Guilt \times Time) experienced due to the experimental manipulation (see Table 2).

A chi-square analysis was performed on each sex organ and sex act mentioned in the fantasies of LSG and HSG subjects. Results indicated that LSG subjects mentioned the words *penis*, *clitoris*, and *sexual intercourse* significantly more often than HSG subjects. No other significant differences between groups were found for sex organs and sex acts that were stated in the fantasies (see Table 3).

HSG subjects checked 30% fewer themes on the Fantasy Checklist as compared to LSG subjects. For each sex guilt group, the number of times a theme was checked was compared to the total number of themes checked. A nonparametric test of the significance between two proportions was performed comparing the percentage of each theme to the total themes for the LSG and HSG groups. LSG subjects responded significantly more often to two themes: "I relive a previous sexual experience" ($z = 2.03$, $p < .05$), and "I have pretended that I am making love to a man that I am acquainted with other than my current lover" ($z = 2.65$, $p < .01$). HSG subjects responded more frequently to two themes related to being dominated sexually: "I imagine that I am being overpowered or forced to surrender," and "I enjoy imagining that I am being dominated sexually and that I am helpless" ($z = 2.78$, $p < .01$, and $z = 2.25$, $p < .05$, respectively). HSG subjects also checked "Thoughts of an imaginary lover or a stranger enter my mind" ($z = 2.17$, $p < .05$), and "I imagine that I am so beautiful that men cannot resist me" ($z = 2.23$, $p < .05$) significantly more often than LSG subjects.

As predicted, ANOVAS on the characteristics of subjects' fantasies resulted in significant differences in response to the conditions of erotic, romantic, or neutral response cues (see Tables 1 and 2). The length of fantasies differed between groups, $F(2, 84) = 4.01$, $p < .05$, and post hoc analyses (see Table 4) revealed that subjects in the erotic fantasy condition wrote longer fantasies than subjects in the romantic condition, $t(58) =$

Table 4
t-Test Results by Condition for Significant F Values

Dependent variable	Erotic vs. romantic	Erotic vs. neutral	Romantic vs. neutral
Length of fantasy	2.68**	1.69	1.23
No. sex organs	3.97**	2.97**	.74
No. sex acts	5.87**	4.14**	1.78
Variety of content	2.25*	1.35	1.16
Sexual arousal	1.95	4.31**	2.38*
Presence of sensations	2.32*	5.28**	2.43*
Genital	2.05*	5.56**	2.78**
Breast	2.82**	4.46**	2.09*
Arousal ^a	1.56	2.95*	1.26

Note. $df = 58$ in all cases.

^a Measured by Multiple Adjective Check List.

* $p < .05$.

** $p < .01$.

2.68, $p < .01$. Both the number of sex acts and sex organs mentioned varied across groups, with subjects in the erotic condition writing more explicit fantasies than those in both the romantic condition and the neutral condition. The variety of content described in each experimental condition produced significant differences, $F(2, 84) = 3.57$, $p < .05$, with post hoc analyses indicating that the differences were a result of the erotic group subjects, who wrote fantasies with more varied content than those in the romantic fantasy group, $t(58) = 2.25$, $p < .05$. Contrary to predictions, the number of fantasies, total number of words written, and number of themes checked were not significantly affected by experimental conditions.

Presence in different experimental groups did not produce differences in the degree of reported guilt, embarrassment, or vividness of the fantasy experience. However, subjects in both the erotic and romantic fantasy groups reported more sexual arousal and the presence of sexual sensations more often than those in the neutral fantasy group. A greater degree of genital and breast sensations were reported by subjects in the erotic group than those in either the romantic or neutral groups, and subjects in the romantic fantasy group reported more genital and breast sensations than those in the neutral group.

As seen in Table 2, none of the predicted interactions between levels of sex guilt and

experimental condition on sexual fantasy were significant.

Discussion

Both sex guilt level in subjects and experimental fantasy cues demonstrated strong effects on fantasy parameters. Sex guilt level had a demonstrable effect on the quantity of fantasy production and females' reports of embarrassment and vividness, whereas the type of fantasy example had more of an effect on the explicitness of fantasies and the report of sexual arousal. The two independent variables did not interact, but they affected different aspects of the sexual fantasy experience for subjects.

Specifically, sex guilt level in females resulted in HSG subjects reporting fewer, shorter, and less explicit fantasies with less variety of content and fewer themes than LSG subjects. These results supported previous research that HSG subjects were more conservative in sexual matters. The findings clearly showed that in explicitness and quantity, sexual fantasy production was under the influence of the sex guilt trait, suggesting that it may be part of a cluster of sexual behaviors governed by this trait.

Experimental fantasy cues exhibited a stronger influence than sex guilt level on explicitness and variety of content of females' fantasies. It is interesting to note that both

HSG and LSG subjects were responsive to the experimental conditions and that the predicted interactions of the effects of the stimuli conditions on HSG and LSG subjects' sexual fantasies did not occur. The speculation that HSG subjects would be more constricted and less explicit in their reported fantasies under the sexually explicit erotic condition as compared to the neutral and romantic conditions was not supported. The strong influence of situational demands on self-reported sexual fantasies for both HSG and LSG subjects suggests that clinicians and researchers need to be sensitive to the conditions under which sexual fantasy production is elicited. Sexual fantasy may be more accurately viewed as being under stimulus control rather than as a stable trait.

Response cuing may, however, place constraints on subjects' responses. The erotic fantasies that the subjects read may not have been explicit enough or varied enough in content to elicit the greatest detail or the maximum number of sexual fantasies that women have experienced. On the other hand, there may be a level of explicitness beyond which subjects would begin to censor their responses. Future research could vary the levels of sexual explicitness in an effort to determine the optimal response facilitator for fantasy production.

As proportionally more HSG females checked themes concerning being sexually dominated and being irresistible to men, it is possible that these themes indicated a reduction in responsibility for the sexual interaction, thus reducing the guilt experienced by these females. The fantasy themes preferred by LSG subjects were more often concerned with real individuals, whereas HSG subjects more often responded to themes concerning an imaginary lover.

Supporting Mosher's (1973) data, sex guilt groups did not differ significantly in degree of self-reported arousal. It is clear from these data that both groups did get at least moderately sexually aroused by sexual fantasy stimulation, and sex guilt did not seem to interfere with reporting this level of arousal. The experimental conditions did produce differences in reported arousal in the predicted direction, with more sexually explicit fantasies (i.e., erotic examples) resulting in

higher levels of arousal. This finding supports Mosher's (1973) results and suggests that sexual arousal in fantasy production seems to be a function of the degree of sexual explicitness of the stimuli rather than a function of the sex guilt trait of individuals. A study assessing actual physiological arousal rather than self-report is needed to more clearly assess the effects of sexual fantasy stimuli on arousal.

HSG subjects reported experiencing more guilt and embarrassment in response to taking part in the experiment and writing sexual fantasies than did LSG subjects, thus supporting previous findings that HSG subjects reported feeling more ashamed after exposure to sexual stimulation. They also indicated that they experienced their fantasies less vividly than did LSG subjects. The question remains whether HSG subjects, in writing fantasies, did not report as many or as explicit fantasies, because they were too embarrassed to do so rather than because they did not experience fantasies in the same way as LSG subjects. However, the more passive response of checking themes also showed the same sex guilt difference. If embarrassment were the important variable affecting fantasy production, one would have expected HSG subjects to write more under the romantic or neutral example condition, in which sexual demand characteristics were minimal. This does not appear to be the case, as HSG subjects, like LSG subjects, wrote more under the sexually explicit condition. Regarding the lower vividness ratings, a reasonable explanation is that HSG subjects, with their postulated censoring mechanisms, probably do not remember their fantasies clearly. As these ratings were initial attempts at assessing parameters of sexual fantasies via self-report, there is a definite need for research that will partial out possible causal factors and their predictive weights in the differential responding of HSG and LSG subjects.

Subjects' ratings in the three experimental conditions, in terms of guilt feelings, embarrassment, and vividness of their fantasies indicated no differential reactions to the fantasy stimuli. These results contradict those of Mosher and Greenberg (1969), who found that subjects responded with more guilt feel-

ings after sexual stimulation in the form of erotic literary passages. Since the most explicit sexual stimuli described only heterosexual intercourse, it is possible that the content of the fantasy examples was not varied enough to elicit feelings of guilt or embarrassment in the subjects. The lack of congruence of these results with the Mosher and Greenberg study suggests the need for further research to manipulate the specific parameters of explicitness and variety of content of experimental stimuli to determine the effects on self-reported guilt levels and ratings of embarrassment.

Most young adult females in this college population related at least one fantasy when asked to write their own sexual fantasies and when prompted with fantasy stimuli. Whether this stimulus situation accurately assessed the incidence of fantasies in females still needs to be determined. Future studies could ask for fantasy incidence information to be collected over a period of time in addition to the self-reported fantasies in a laboratory setting.

References

- Abraham, K. Manifestations of the female castration complex. *International Journal of Psychoanalysis*, 1922, 3, 1-29.
- Eidelberger, L. A contribution to the study of the masturbation phantasy. *International Journal of Psychoanalysis*, 1945, 26, 127-137.
- Evans, D. R. Masturbatory fantasy and sexual deviation. *Behaviour Research and Therapy*, 1968, 6, 17-19.
- Friday, N. *My secret garden: Women's sexual fantasies*. New York: Trident Press, 1973.
- Galbraith, G. G., & Mosher, D. L. Associative sexual responses in relation to sexual arousal, guilt, and external approval contingencies. *Journal of Personality and Social Psychology*, 1968, 10, 142-147.
- Hariton, E. B., & Singer, J. L. Women's fantasies during sexual intercourse. *Journal of Consulting and Clinical Psychology*, 1974, 42, 313-322.
- Hollender, N. H. Women's coital fantasies. *Medical Aspects of Human Sexuality*, 1970, 4, 63-70.
- Kaplan, H. K. *The new sex therapy*. New York: Brunner/Mazel, 1974.
- Kinsey, A. C., Pomeroy, W. B., Martin, C. E., & Gebhard, P. H. *Sexual behavior in the human female*. Simon & Schuster: New York, 1953.
- Kline-Graber, G., & Graber, B. *Woman's orgasm*. Indianapolis, Ind.: Bobbs-Merrill, 1975.
- Langston, R. D. Sex guilt and sex behavior in college students. *Journal of Personality Assessment*, 1973, 37, 467-472.
- Marquis, J. Orgasmic reconditioning: Changing sexual object choice through controlling masturbation fantasies. *Journal of Behavior Therapy and Psychiatry*, 1970, 1, 263-271.
- Marshall, W. L. The modification of sexual fantasies. *Behaviour Research and Therapy*, 1973, 11, 557-564.
- Mosher, D. L. Interaction of fear and guilt in inhibiting unacceptable behavior. *Journal of Consulting Psychology*, 1965, 29, 161-167.
- Mosher, D. L. The development of a multitrait-multimethod matrix analysis of three measures of three aspects of guilt. *Journal of Consulting Psychology*, 1966, 30, 25-29.
- Mosher, D. L. Sex differences, sex experience, sex guilt, and explicitly sexual films. *Journal of Social Issues*, 1973, 29, 95-112.
- Mosher, D. L., & Cross, H. J. Sex guilt and premarital sexual experiences of college students. *Journal of Consulting and Clinical Psychology*, 1971, 36, 27-32.
- Mosher, D. L., & Greenberg, I. Females' affective responses to reading erotic literature. *Journal of Consulting and Clinical Psychology*, 1969, 33, 472-477.
- Mussen, P. H., & Scodel, A. The effects of sexual stimulation under varying conditions on TAT sexual responsiveness. *Journal of Consulting Psychology*, 1955, 19, 90.
- Nowlis, V. Research with the Mood Adjective Check List. In S. C. Tompkins & C. E. Izard (Eds.), *Affect, cognition and personality*. Springer: New York, 1965.
- Okel, E., & Mosher, D. L. Changes in affective states as a function of guilt over aggressive behavior. *Journal of Consulting Psychology*, 1968, 32, 265-270.
- Ray, R. E., & Walker, C. E. Biographical and self-report correlates of female guilt responses to visual erotic stimuli. *Journal of Consulting and Clinical Psychology*, 1973, 41, 93-96.
- Routh, D. R., Warehime, R. G., Gresen, R., & Roger, L. Measuring the ability to express sexual thoughts. *Journal of Personality Assessment*, 1973, 37, 342-350.
- Schill, T. Need for approval, guilt, and sexual stimulation, and their relationship to sexual responsibility. *Journal of Consulting and Clinical Psychology*, 1972, 38, 31-35.
- Schill, T., & Chapin, J. Sex guilt and males' preferences for reading erotic magazines. *Journal of Consulting and Clinical Psychology*, 1972, 39, 516.
- Schmidt, G. Male-female differences in sexual arousal and behavior during and after exposure to sexually explicit stimuli. *Archives of Sexual Behavior*, 1975, 4, 353-365.
- Schwartz, S. Effects of sex guilt and sexual arousal on the retention of birth control information. *Journal of Consulting and Clinical Psychology*, 1973, 41, 61-64.
- Sterba, R. Dreams and acting out. *Psychoanalytic Quarterly*, 1921, 2, 175-180.
- Wulff, M. A case of male homosexuality. *International Journal of Psychoanalysis*, 1942, 23, 112-120.

On the External Validity of Two Psychotherapy Analogues

Kenneth Kushner
University of Michigan

The present study investigated the similarity between two live psychotherapy analogues and real psychotherapeutic interviews. Therapists were asked to participate in two different types of analogue situations and in initial intake sessions with real clients. In both of the analogues, a recruited subject presented a real personal problem to the therapists in helping interactions. Audiotapes of the real and analogue interviews were rated on 10 dependent variables, which were different dimensions of therapist and client behaviors. Different results were obtained for each analogue. The major findings concern mean differences between the analogue and real interviews and the linear relationships between the real and analogue interviews. Additional findings, including significant interactions between the type of interview and the experience level of the therapist, are also discussed. The results are interpreted as indicating that the generalizability of the analogues is contingent on the dependent variables in question, the type of relationship to be predicted, and the experience level of the therapists. The implications that these results have for future research involving psychotherapy analogues are discussed.

The present research was inspired by the importance of analogue methodologies for process studies of psychotherapy. Since the time of the first study that used an analogue of the psychotherapeutic process (Keet, 1948), the use of psychotherapy analogues (also referred to as simplifications or simulations) has become a very popular and accepted research strategy. The types of analogues that researchers have used over the years have been both diverse and creative. They have ranged from highly artificial, non-live situations (i.e., Porter, 1950) to highly realistic live situations (i.e., Russell & Snyder, 1963). Today, the sheer number of ana-

logue studies is large enough to warrant extensive literature reviews (Heller, 1971; Munley, 1974), as well as substantive theoretical critiques (Bordin, 1965; Strong, 1971; Strupp, 1961; Thomas, 1962). It is easy to understand why analogues are often preferable to the naturalistic study of psychotherapy; they are more convenient, they afford more experimental control, and they avoid many of the ethical problems inherent in field study methodologies that require the use of real clients and therapists.

There is an underlying assumption involved in the use of analogues in psychotherapy research; namely, that the results found in the simplified situations are generalizable to the real psychotherapeutic setting that they are supposed to simulate. As Heller and Marlatt (1969) have pointed out, this is a question of the external validity of analogues. The utility of an analogue depends on the amount of confidence one can place in one's ability to extrapolate from results found in the analogue situation to the real psychotherapeutic situation to which it refers. The degree of similarity between results found in analogue and naturalistic settings is an empirical ques-

This article is based on a doctoral dissertation undertaken at the University of Michigan under the chair of Edward S. Bordin. A brief version of this report was presented at the Eighth annual meeting of the Society for Psychotherapy Research, Madison, Wisconsin, June 1977.

The author would like to thank E. S. Bordin and Erica Serlin for their thoughtful suggestions regarding the present article.

Requests for reprints should be sent to Kenneth Kushner, who is now at the Department of Family Medicine, Medical College of Ohio, C.S. # 10008, Toledo, Ohio 43699.

tion. Bordin (1965) advocated validation studies that "bridge" results found in the one setting to the other. Other authors have also suggested such empirical comparisons (e.g., Strupp, 1961). Nevertheless, as of today, validation studies can be counted on one hand (Hopke, 1955; Matarazzo, Wiens, Matarazzo, & Saslow, 1968; Roark, 1969; Sigal, Guttman, Chagoya, & Lasry, 1975; Sigal, Lasry, Guttman, Chagoya, & Pilon, 1977). In each of these five studies, the investigators found some degree of similarity between the analogue and real interviews for some of the dependent variables that they examined. However, for other dependent variables, the investigators found important differences between real and analogue situations. These results indicate that there is good reason not to accept on faith the generalizability of all results based on analogues.

The present study was undertaken to apply the principal of "empirical bridging" to two similar psychotherapy analogues. In both analogues, real therapists interviewed nonreal, or "quasi," clients in simulated initial intake sessions. In one type of analogue, the *nonstandard client analogue* (similar to Danish, D'Augelli, & Brock, 1976), each therapist interviewed a different female volunteer drawn from a subject pool. The subject was instructed to genuinely discuss a personal problem with the therapist. In the second analogue, the *standard client analogue* (similar to Carkhuff, Kratochvil, & Friel, 1968), all therapists in the sample interviewed the same female quasi-client. She was instructed to present the same real problem to each therapist and to discuss it genuinely. The therapists were instructed to conduct both types of analogues as if they were real initial intake interviews. Thus, the two types of analogues were identical, except for the fact that one interviewee was used in the standard client analogue and multiple interviewees were used in the nonstandard client analogue.

Method

Design

The present study contained a within-subjects design, in that each therapist participated in the standard client analogue, the nonstandard client analogue, and the real initial intake interviews. All

interviewees were female, and the order of presentation of the two analogues was balanced in the sample. This design allowed for direct comparison of the process of each analogue to the real therapy interviews.

Subjects and Procedure

Therapists were recruited from four campus mental health agencies. Thirty-one therapists volunteered to participate in the study. The minimum qualifications for therapists to be included in the data analysis was having completed at least one real and one analogue interview. As a result, 23 therapists were included in the data. Of these, 15 therapists participated in both analogues and two real interviews, 7 therapists participated in both analogues but were able to record only one real intake interview, and 1 therapist participated in only the nonstandard client analogue but recorded two real initial intake interviews. The remaining 8 therapists, those who volunteered to participate but who did not record at least one real intake session, all reported that they were unable to obtain appropriate intake interviews. The therapists who participated in the study were a diversified group, as shown by an even sex distribution and wide ranges of age, experience level, and theoretical orientation.

The clients for the *real therapy* interviews were selected by the therapists. Each therapist was asked to tape record the first two real intake sessions that met the following criteria: (a) The client had to be female; (b) the presenting problem had to be emotional rather than solely educational or vocational in nature; (c) it had to last at least 30 minutes; and (d) the client had to agree to participate in the research.

The subjects in the nonstandard client analogue interviews were females recruited by phone from a paid subject pool. They were told that the study would entail having an interview with a therapist in which they would talk genuinely about a real personal concern or problem. The only prerequisite was that they were not currently involved in counseling or psychotherapy.

The client for the *standard client analogue* interview was a 21-year-old senior undergraduate psychology major. She was instructed to choose a real personal problem, and other than presenting it in the same words to each therapist, she was asked to discuss it genuinely and spontaneously in each interview. Thus, her instructional set was essentially the same as that of the nonstandard clients. The standard presenting problem that she chose concerned her ambivalence about her relationship with her boyfriend, especially whether she should move to the east coast when she graduated instead of staying in the midwest to be near him. She described the problem on the information sheets as follows for each interview: "Confusion about various things—i.e., boyfriend and where I go at the end of the year—how can I make some important decisions related to above—do I think too much about it, etc."

In designing the two analogues, the attempt was made to create situations that differed from the real therapy interviews only by virtue of the fact that the analogue subjects were actively recruited to discuss real problems, whereas the real therapy clients were self-motivated to seek help for their problems. For this reason, all interviews (analogue and real) were conducted in the therapists' own offices. In addition, the analogue subjects, as well as the real clients, presented to the therapist a face sheet providing both demographic information and their presenting problem(s). In both analogue interviews, the instructional set was designed to parallel the demand characteristics of a real therapy interview; the subjects were requested to present real problems, and the therapists were asked to conduct the interviews as if they were actual initial intakes. The therapists were aware of which real clients were being taped for research purposes. They were also informed that the analogue subjects were experimental subjects. However, they were not told that there were any differences between the two analogue interviews. Specifically, an attempt was made to keep the therapists blind to the fact that one of the subjects was being seen by other therapists. During debriefing, eight therapists reported knowledge or suspicion that a standard client was used.

It is worth noting that I judged the real and analogue situations to be equivalent in terms of one more important characteristic—presenting problem. The presenting problems of the real and nonstandard clients mostly centered on interpersonal and/or academic concerns. Difficulties with parents, boyfriends, or schoolwork were typical examples. In light of this fact, the standard client's presenting complaint was representative of the type of problems presented by the other subjects. It is also worth noting that the mean age of the real clients was 22.3 and that of the nonstandard clients was 19.9. This difference was significant, $t(59) = 3.33$, $p < .01$. Although the attempt was made to keep the real clients, standard client, and nonstandard client interviews equivalent in length by asking the participants to keep them between $\frac{1}{2}$ and 1 hour, their mean lengths were 49 min 3 sec, 40 min 22 sec, and 42 min 49 sec, respectively. Only the contrast between the real client and standard client analogue interviews was significant ($p < .05$).

Dependent Variables

The audiotape recordings of the real client, standard client, and nonstandard client analogue interviews were rated for the following 10 dependent variables that were selected due to their relevance to psychotherapy research in general and analogue research specifically: accurate empathy, nonpossessive warmth, genuineness (all after Truax & Carkhuff, 1967), ambiguity (after Osburn, 1951), question-to-statement ratio (Q/S; after Ornston, Cicchetti, Levine, & Fierman, 1968), mean duration of therapist utterance (TUL), mean duration of client utterance (CUL), percentage of client talk time (C%), percent-

age of therapist talk time (T%), and percentage of silence (S%; all after Matarazzo et al., 1968).

For accurate empathy, nonpossessive warmth, genuineness, and ambiguity, the ratings were done on four excerpts per tape, each approximately 4 minutes in duration. This yielded four scores per interview. A mean score for each variable was then calculated for each interview. Unless stated otherwise, it was these mean scores that were used in the data analysis. The remaining dependent variables—Q/S, TUL, CUL, C%, T%, and S%—were rated for the entire tape by a hand-held stopwatch, yielding one overall score per interview per variable. For one therapist whose data was included in the analysis, the excerpts were inaudible, and the raters were unable to rate her interviews for empathy, warmth, genuineness, and ambiguity. However, they were able to rate the original tapes for the remaining variables.

Raters and Rater Reliability

The raters were six undergraduate psychology majors. Three of them rated the interviews for accurate empathy, nonpossessive warmth, and genuineness. The other three rated them for the remaining dependent variables. Both groups of raters were given extensive training and practice in the use of the scales by me, and each group had to demonstrate a high degree of interrater reliability before the actual rating began. For both groups of raters, prerating reliability was determined on the basis of their scores on 11 4-minute excerpts taken from therapy sessions that were not in the present sample. In addition to the prerating reliabilities, each group of raters rated the same 11 excerpts spread over the actual ratings. These scores allowed me to determine the "drift" in reliability that might have occurred in the process of rating. In determining reliability, correlation coefficients were computed for each of the three pairs of raters within each group, and a mean of these three coefficients was then calculated. Instead of calculating the coefficient for the Q/S ratio, separate coefficients were calculated for the number of questions and the number statements.

The prerating reliabilities were uniformly high, ranging from correlations of .75 for warmth to .99 for CUL. The drift rating coefficients were also high for all dependent variables except empathy and warmth, which were .01 and .53, respectively. Since this raises questions about the value of the ratings for empathy and warmth, the results pertaining to these two variables have been omitted from the following presentation. Correlations for all prerating and drift rating reliabilities can be found in Kushner (1977).

Results

Mean Differences Between Analogue and Real Interviews and Effects of Therapist Experience Level

To investigate whether there were mean differences between the real and analogue

Table 1
Cell Means for Conditions Main Effect for Both Analogues Versus Real Interviews

Variable	Condition			<i>F</i>	<i>df</i>
	Real	Standard	Nonstandard		
Genuineness	2.31	2.36	2.19	1.23	2,28
Ambiguity ^a	13.98	19.22	15.75	2.61	2,28
CUL (in sec)	20.51	10.64	19.78	5.67*	2,32
TUL (in sec)	6.97	8.20	7.63	3.22	2,32
Q/S	1.72	1.87	1.38	2.04	2,32
C% ^b	69.73	56.93	69.17	9.98***	2,32
T% ^b	22.06	29.84	23.08	3.72*	2,32
S% ^b	8.47	13.22	7.88	3.10	2,32

Note. CUL = mean duration of client utterance; TUL = mean duration of therapist utterance; Q/S = question to statement ratio; C% = percentage of client talk time; T% = percentage of therapist talk time; S% = percentage of silence. For Dunnett's test, real vs. standard analogue on CUL and C%, $p < .001$.

^a Low score indicates more ambiguity.

^b Analysis of variance performed on arc-sin transformations of raw scores. Cell means are expressed as means of raw scores before transformation.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

interviews, an analysis of variance of the form Conditions \times Experience \times Therapists was calculated for each dependent variable. There were three levels of the conditions factor (real, standard, and nonstandard).¹ Experience was included as a factor due to the large range of clinical experience in the sample of therapists (0–20 years) and the fact that it was wondered whether the analogues would reflect differences between less experienced and more experienced therapists. Accordingly, the sample was split at the median into a group of less experienced (0–3 years) and more experienced (5+ years) therapists. This yielded two levels of the experience factor. Thus, therapists were nested within experience, and conditions were crossed with both experience and therapists. Several therapists had to be randomly discarded in order to balance the cells, resulting in 16 therapists in those analyses conducted on genuineness and ambiguity and 18 in those conducted on the remaining variables. For those variables that were bounded ratios (C%, T%, and S%), the analyses were conducted on the arc-sin transformations of the raw scores.

The cell means and *F* ratios of the conditions main effect are displayed in Table 1.² To determine whether the real interviews differed from either analogue, Dunnett's method

of testing all cell means against one control (Winer, 1971) was used to test the significance of the two pairwise contrasts of interest: between the real and standard client interview conditions and between the real and nonstandard client interview conditions. The results of Dunnett's test for these variables that had significant overall *F* ratios appear in Table 1. As can be seen, the conditions main effect was significant for CUL and C%. For these two client variables, the contrasts between the real client and standard client conditions were significant, revealing that the standard client had shorter mean durations of utterance length and talked for smaller percentages of the sessions than did the real

¹ In all analyses of variance, the scores of the first real interviews were used for those therapists who had recorded two real initial intake sessions. Prior to conducting the analyses, *t* tests had shown that there were no significant differences between the first and second real interviews.

² Actually the analyses for genuineness and ambiguity were conducted as repeated measures analyses of variance. This was because four ratings of those variables were made from each tape, one from each excerpt. Thus, the design for these variables was Conditions \times Experience \times Therapists \times Replicates; with 3, 2, 8, and 4 levels of each factor, respectively. (For a further treatment of this, see Kushner, 1977.)

Table 2
Cell Means for the Experience Main Effect

Variable	Less experienced (≤ 3 yr)	More experienced (≥ 5 yr)	<i>F</i>	<i>df</i>
Genuineness	2.15	2.43	3.45	1,14
Ambiguity ^a	13.33	19.30	7.93**	1,14
CUL (in sec)	22.02	11.94	7.39**	1,16
TUL (in sec)	6.51	8.69	3.22	1,16
Q/S	2.20	1.11	6.21*	1,16
C% ^b	70.26	60.29	9.06**	1,16
T% ^b	20.14	29.85	7.56**	1,16
S% ^b	9.59	10.13	.12	1,16

Note. CUL = mean duration of client utterance; TUL = mean duration of therapist utterance; Q/S = question to statement ratio; C% = percentage of client talk time; T% = percentage of therapist talk time; S% = percentage of silence.

^a Low score indicates more ambiguity.

^b Analysis of variance performed on arc-sin transformations of raw scores. Cell means presented are means of raw scores before transformation.

* $p < .05$.

** $p < .01$.

clients. For the third variable, T%, the contrast between the real client and standard client conditions approached but did not attain

Table 3
Cell Means for Experience \times Conditions
Interaction for Ambiguity, C%, and T%

Condition	Experience	
	Less	More
Ambiguity ^a		
Real	7.97	20.00
Standard	19.20	19.25
Nonstandard	12.83	18.67
C% ^b		
Real	78.26	61.20
Standard	57.30	56.57
Nonstandard	75.24	63.09
T% ^c		
Real	13.18	30.95
Standard	29.99	29.70
Nonstandard	17.27	28.90

Note. C% = percentage of client talk time; T% = percentage of therapist talk time. Analysis of variance was performed on arc-sin transformation of raw scores. Cell means are means of raw scores before transformation.

^a Low score indicates more ambiguity; $F(2, 28) = 3.29$, $p < .05$.

^b $F(2, 32) = 3.48$, $p < .03$.

^c $F(2, 32) = 4.14$, $p < .03$.

significance.³ In no instance was the contrast between the real client and the non-standard client interviews significant.

Table 2 displays the cell means and *F* ratios for the experience main effects. As can be seen, there were significant differences between the less experienced and more experienced therapists for the following dependent variables: ambiguity, CUL, Q/S, C%, and T%. Significant Experience \times Conditions interactions were found for three dependent variables ambiguity, C%, and T%. The cell means and *F* ratios of these significant interactions are displayed in Table 3. The same two patterns are reflected in the cell means of all three of those interactions. The first pattern is that the differences between the less and

³ A second set of analyses of variance, based on a larger sample size, did in fact reveal a significant difference between the real and standard client conditions for T%. This was when the analyses were calculated without the experience factor (Conditions \times Experience). This enabled the inclusion of the data of three more therapists. (They had originally been disregarded in order to balance the cells for the previous series of analyses.) Similarly, this second set of analyses revealed a significant conditions main effect for S%, $F(2, 42) = 3.38$, $p < .05$, and a significant contrast between the real and standard client analogue conditions for that variable ($p < .05$). These results, and the summaries of all the analyses, are discussed in Kushner (1977).

more experienced therapists that were evident in the experience main effect were found in the real client and nonstandard client interviews but not in the standard client interviews. Rather, the cell means for the standard client interviews were almost identical for the less experienced and more experienced therapists.

The second pattern that was reflected in the cell means of the significant Experience \times Conditions interactions for ambiguity, C%, and T% is that there was much greater range in the cell means across conditions for the interviews conducted by the less experienced therapists than there were for the more experienced therapists. For the less experienced therapists, the disparities between the means of the real client and standard client conditions were the most pronounced. For the more experienced therapists, the cell means of all three conditions seemed roughly equivalent. Apparently, the former therapists were more influenced by the differences in the type of subjects that they were interviewing than were the latter therapists.

Linear Relationships Between the Real and Analogue Interviews

The linear relationships between the real and analogue interviews address the ability to make predictions regarding the relative performance of specific therapists in real therapy on the basis of his/her performance in an analogue. This was investigated via two sets of correlations. One was based on the same therapists' scores in the real interviews and in the standard client analogue; the other was based on the same therapists' scores in the real interviews and in the nonstandard client analogues. The correlations were calculated using the mean of the first and second real interviews for those therapists who participated in two initial intake sessions. The results of the correlations are summarized in Table 4. An analogue can be said to be a predictor of real therapy behavior if the coefficient is positive and significant. The standard client analogue was a predictor for three dependent variables: genuineness, CUL, and Q/S. The nonstandard client analogue

Table 4
Correlations of Scores of Same Therapist's Real Interviews Versus Each Analogue Interview

Variable	Comparison			
	Real ^a vs. standard		Real ^a vs. nonstandard	
	<i>r</i>	<i>df</i>	<i>r</i>	<i>df</i>
Genuineness	.77*	19	.20	20
Ambiguity	.27	19	.10	20
CUL	.56**	20	.30	21
TUL	.12	20	.21	21
Q/S	.65**	20	.54**	21
C%	.19	20	.23	21
T%	.08	20	.35	21
S%	.23	20	-.42*	21

Note. CUL = mean duration of client utterance; TUL = mean duration of therapist utterance; Q/S = question to statement ratio; C% = percentage of client talk time; T% = percentage of therapist talk time; S% = percentage of silence.

^a Means of raw scores of first and second initial interviews were used for those therapists who submitted two real interviews.

* $p < .05$.

** $p < .01$.

was a predictor of real therapy behavior only for Q/S.

A supplemental set of analyses revealed an interesting and somewhat surprising result. The correlations were calculated between the first and second real interviews for those therapists who recorded two initial intake sessions. The coefficients were quite low, with none of them attaining statistical significance. (See Kushner, 1977, for coefficients.) This indicates that behavior in one initial intake session did not adequately predict behavior in a second initial intake that the same therapist conducted with another client. It should be noted that the correlations between the two real interviews did not differ statistically from either the real and standard client analogues or the real and nonstandard client analogues. Thus it cannot be concluded that a second real interview is a better or worse predictor of a therapist's real interview than either the standard client or nonstandard client analogue.

Table 5

Variances of Dependent Variables with F Tests for Pairwise Comparisons Against Standard Client Condition

Variable	Variance			Standard vs. real		Standard vs. nonstandard	
	Standard	Real	Nonstandard	F	df	F	df
Genuineness	.195	.161	.091	1.17	20,21	2.14	20,21
Ambiguity	40.208	66.171	80.534	1.92	21,20	2.00	21,20
CUL	15.234	276.590	264.323	7.38*	22,21	17.34*	22,21
TUL	8.951	8.861	20.917	1.15	21,22	2.34	22,21
Q/S	1.100	1.442	.806	1.88	22,21	.136	21,22
C% ^a	.009	.033	.458	3.94*	22,21	5.06*	22,21
T% ^a	.015	.014	.018	1.15	21,22	1.94	22,21
S% ^a	.007	.004	.006	1.46	21,22	1.16	22,21

Note. CUL = mean duration of client utterance; TUL = mean duration of therapist utterance; Q/S = question to statement ratio; C% = percentage of client talk time; T% = percentage of therapist talk time; S% = percentage of silence.

^a Calculations of variance conducted on arc sin transformations of raw scores.

* $p < .01$.

Differences in Variance Between the Standard Client and Other Interview Conditions

The last set of analyses does not relate to the generalizability of the analogues, but rather to the rationale behind the use of a standard client. The single interviewee should theoretically provide a more constant stimulus than multiple interviewees. Therefore, there should be less variance for the dependent measures in the standard client condition than in the real client or nonstandard client conditions. This issue was addressed by calculating the variance in each condition and then conducting two *F* tests—one contrasting the dependent measures for the standard client analogue and real interviews and the other contrasting the standard client and the nonstandard client interviews. The variables and *F* tests are shown in Table 5. The *F*s were significant only for CUL and C%, showing that there was considerably less variance in client behavior for these variables in the standard client interviews than in either the real client or nonstandard client interviews.

Discussion

Several main findings regarding the similarity between the analogue and real interviews were obtained in the present study. These will be discussed with reference to

their implications regarding the future use of the analogues: First, there were no mean differences between the real client interviews and the nonstandard client analogues for any of the eight dependent variables. However, there were significant mean differences between the real client interviews and the standard client analogues for mean duration of client utterance and mean duration of therapist utterance. These mean differences indicate that the levels of functioning of the standard client would not be predictive of the magnitude of the means of the real therapy clients. The significant Conditions \times Experience interactions for ambiguity, C%, and T%, which showed greater ranges in the cell means across conditions for the less, but not the more, experienced therapists indicate that there were mean differences between the real interviews and the standard client analogue for the former therapists only. These significant Conditions \times Experience interactions also indicate that the standard client analogue did not reflect the mean differences for ambiguity, C%, and T% found in the real interviews between the less and more experienced therapists, but the nonstandard client analogue would have predicted those differences. For the remaining dependent variables, the lack of significant Conditions \times Experience interactions indicates the abil-

ity of both analogues to reflect true differences (or lack of differences) between interviews conducted by less experienced and more experienced therapists.

The correlations between the real and analogue interviews addressed the ability to predict the relative performance of individual therapists in real therapy on the basis of their behavior in the analogues. The analogues were rather disappointing as predictors. The standard client analogue was found to be a predictor of genuineness, CUL, and Q/S. The nonstandard client analogue was found to be a predictor for Q/S only. Although the low correlations between the two real interviews might be attributable to the small sample size on which they based ($ns = 14$ and 16), other authors (Beutler, Neville, & Workman, 1973) have reported low correlations between real interviews of the same therapists. This implies that therapists might be highly variable from session to session.

It should be clear from the above summary that the external validity of an analogue is not an all-or-nothing property; rather, the utility of an analogue as a predictor of real therapy behavior may be dependent on several factors, such as (a) the dependent variables in question, (b) the type of relationship that one would want to predict (i.e., mean differences between groups of therapists or the relative performance of individual therapists), and (c) the experience of the therapists participating in the analogue. It should be noted that the present study cannot be seen as a blanket validation of the use of the two analogues. Rather, conclusions about the validity of any analogue should be specific, in that they should define the variables and conditions for which it is concluded to be predictive of real therapy behaviors. The onus is on the researcher to empirically establish continuity of style between the analogue and real therapy, rather than assuming it to be there for all variables.

It is obvious that the results obtained for the standard client in the present study may have been ideosyncratic of her particular interviewing style (characterized by relatively less verbal behavior and shorter interview lengths) and thus may not be general-

izable to other standard clients. A study that contrasted the interviews of several standard clients to real interviews with the same therapists would address this issue.

In addition to the study proposed above, the present study has several other implications for future research: (a) It adds more evidence that results found in analogues may not be generalizable to real therapy behaviors. Therefore, more validation studies should be undertaken for other analogue situations. (b) Interpretation of results based on nonvalidated analogues should be tentative in nature. (c) Naturalistic studies appear to be largely preferable to analogue studies from the standpoint of rigor. (d) The consistency of therapists in real interviews should continue to be studied. If it is true that therapists are highly variable from interview to interview, then sampling several interviews by the same therapist would be a more valid procedure for future research than merely sampling one interview.

References

- Beutler, D. T., Neville, C. W., & Workman, S. N. Some sources of variance in "accurate empathy" ratings. *Journal of Consulting and Clinical Psychology*, 1973, 40, 167-169.
- Bordin, E. S. Simplification as a research strategy in psychotherapy. *Journal of Consulting Psychology*, 1965, 29, 493-503.
- Carkhuff, R. R., Kratochvil, D., & Friel, T. The effects of professional training: The communication and discrimination of facilitative conditions. *Journal of Counseling Psychology*, 1968, 15, 68-74.
- Danish, S. J., D'Augelli, A. R., & Brock, G. W. An evaluation of helping skills training effects on helper's verbal responses. *Journal of Counseling Psychology*, 1976, 23, 259-266.
- Heller, K. Laboratory interview research as an analogue to treatment. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change*. New York: Wiley, 1971.
- Heller, K., & Marlatt, G. A. Verbal conditioning, behavior therapy and behavior change: Some problems in extrapolation. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- Hopke, W. E. The measurement of counselor attitudes. *Journal of Counseling Psychology*, 1955, 2, 212-216.
- Keet, C. D. Two verbal techniques in a miniature counseling situation. *Psychological Monographs*, 1948, 62(7, Whole No. 294).
- Kushner, K. P. On the external validity of two psychotherapy analogues (Doctoral dissertation, Uni-

- versity of Michigan, 1977). *Dissertation Abstracts International*, 1977, 38, 2867b. (University Microfilms No. 77-26, 286)
- Matarazzo, J. D., Wiens, A. N., Matarazzo, R. G., & Saslow, G. Speech and silence behavior in clinical psychotherapy and laboratory correlates. In J. M. Shlien (Ed.), *Research in psychotherapy* (Vol. 3). Washington, D.C.: American Psychological Association, 1968.
- Munley, P. H. A review of counseling analogue research methods. *Journal of Counseling Psychology*, 1974, 21, 320-331.
- Osburn, H. G. An investigation of the ambiguity of counselor behavior. (Doctoral dissertation. University of Michigan, 1951). *Dissertation Abstracts*, 1951, 221. (University Microfilms No. 00-03,544)
- Ornston, P. S., Cicchetti, D. V., Levine, J., & Fierman, L. B. Some parameters of verbal behavior that reliably differentiate novice from experienced psychotherapists. *Journal of Abnormal Psychology*, 1968, 73, 240-244.
- Porter, E. H. *An introduction to therapeutic counseling*. Cambridge, Mass.: Riverside Press, 1950.
- Roark, A. E. The influences of training on counselor responses in actual and role-playing interviews. *Counselor Education and Supervision*, 1969, 8, 289-295.
- Russell, P. D., & Snyder, W. U. Counselor anxiety in relation to amount of clinical experiences and quality of affect demonstrated by clients. *Journal of Consulting Psychology*, 1963, 27, 358-363.
- Sigal, J. J., Guttman, H., Chagoya, L., & Lasry, J. C. Predictability of family therapists' behavior. *Canadian Psychiatric Association Journal*, 1975, 18, 199-202.
- Sigal, J. J., Lasry, J. C., Guttman, H., Chagoya, L., & Pilon, R. Some stable characteristics of family therapists' intervention in real and simulated therapy. *Journal of Consulting and Clinical Psychology*, 1977, 45, 23-27.
- Strong, S. R. Experimental laboratory research in counseling. *Journal of Counseling Psychology*, 1971, 18, 106-110.
- Strupp, H. H. The therapist's contribution to the research process: Beginnings and vagaries of a research program. In H. H. Strupp & L. Luborsky (Eds.), *Research in psychotherapy*. (Vol. 2). Washington, D.C.: American Psychological Association, 1961.
- Thomas, E. J. Experimental analogues of the case-work interview. *Social Work*, 1962, 7, 24-30.
- Truax, C. B., & Carkhuff, R. R. *Toward effective counseling and psychotherapy: Training and practice*. Chicago: Aldine, 1967.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

Received November 28, 1977 ■

Are Special Norms for Minorities Needed? Development of an MMPI *F* Scale for Blacks

Malcolm D. Gynther
Auburn University

David Lachar
Lafayette Clinic
Detroit, Michigan

W. Grant Dahlstrom
University of North Carolina

As part of a large-scale investigation of the need for special Minnesota Multiphasic Personality Inventory (MMPI) norms for black adult American test subjects, 882 normal black adults from Alabama, Michigan, and North Carolina were tested. Item analyses of these protocols revealed 33 items that met the 10% or less endorsement criterion used to develop the MMPI *F* scale. Comparisons were made between white and black endorsements of the items on this new scale, items of the standard *F* scale, and additional items in the MMPI pool that met the 10% criterion but were not included in the original *F* scale. White endorsement patterns agreed with the blacks' *F* scale, but black endorsement patterns agreed with only one third of the standard *F*-scale items. Further, black adults showed comparable levels of infrequency on only 6 of the 38 supplementary *F* items. Although these results do not necessarily indicate that special norms for the clinical scales are necessary, the amount of difference between responses of blacks and whites to rarely endorsed items suggests that for blacks, this new scale may be a more accurate measure of correlates associated with endorsement of deviant items than the standard *F* scale.

The *F* scale of the Minnesota Multiphasic Personality Inventory (MMPI) was proposed as one of the validity indicators by the test's authors, Hathaway and McKinley (1951). The 64 items comprising this scale were selected in part on the basis of the frequency of endorsement and in part on the basis of diversity of item content. That is, the items chosen were answered in the scored direction by 10% or less of the subjects in the Minnesota normative adult samples, and they covered a wide range of topics, with only a few items referring to any one area of behavior or experience. The item tallies used

to identify potential *F*-scale items came from an early subsample of the Minnesota normal subjects (111 men and 118 women); subsequent analyses reported in Hathaway and Briggs (1957) and Dahlstrom, Welsh, and Dahlstrom (1975) indicate that on the more complete data of the Minnesota samples, 3 of the 64 items do not meet the 10% or below criterion, and 1 additional item satisfies this criterion only for females. In addition, there are 38 items that could have been included in the *F* scale, but they were excluded by the test's authors even though they were endorsed by 10% or less of the revised Minnesota normal adult group.

High scores on this scale (typically $F > 16$) signified that the profile was invalid due to the subject's carelessness or lack of comprehension. Other reasons for obtaining elevated scores were recognized early in the development of the inventory. Meehl and Hathaway (1946) stated that schizoid subjects

This investigation was supported in part by National Institute on Drug Abuse Grant 7H81 DA-01879-01. The authors wish to thank James L. Grisell, Michael Butkus, Roger Gudobba, and Paul Sherwood for assistance in data analysis.

Requests for reprints should be sent to Malcolm D. Gynther, Department of Psychology, Auburn University, Auburn, Alabama 36830.

and subjects who apparently wished to put themselves in a bad light also obtained high scores on this scale. Dahlstrom, Welsh, and Dahlstrom (1972) have described several other possible reasons for high scores, such as random responding, answering true to all items, cyclical alternations of true and false answers, a cry for help or pleading for special attention, a bilingual background with English as a second language, visual impairment, and acute psychotic disorganization.

Research has shown personological and psychopathological correlates of elevated MMPI *F* scores obtained by whites. Among persons relatively free of serious problems in living, the Institute of Personality Assessment and Research studies (Gough, McKee, & Yandell, Note 1) characterized those with moderately high *F* scale scores as moody, restless, dissatisfied, and opinionated. Carson (1969) has suggested that *T* scores in the range of 65 to 80 frequently appear in sullen, rebellious personalities of the schizoid, antisocial, or Bohemian type. High MMPI *F* scores obtained from court cases referred for diagnostic evaluation have been shown to be related to a diagnosis of psychopathy (Gynther, 1961; Gynther & Shimkunas, 1965), as well as to the commission of serious sex crimes (Gynther, 1962). Elevated *F* scores obtained from psychiatric patients appear to be associated with withdrawal, poor judgment, short attention span, delusions, and hallucinations (Gynther, Altman, & Warbin, 1973). Other correlates of moderate to high MMPI *F* scores can be found in Lachar (1974), Duckworth and Duckworth (1975), and Graham (1977). Dahlstrom et al. (1975) have summarized the research (provided that the sources of protocol invalidity mentioned earlier can be dismissed) as indicating that "the degree of emotional disturbance in the individual can be judged with reasonable accuracy from the elevation of the *F* scale" (p. 31).

It is interesting to note, in this context, that black subjects typically obtain higher scores on the *F* scale of the MMPI than white subjects. There have been 10 studies to date comparing noninstitutionalized blacks and whites on the MMPI. In all but 1 of

these studies (several of which attempted to control for social class), blacks obtained significantly higher scores than whites on this scale. Fifteen studies have compared black and white institutionalized subjects (i.e., psychiatric inpatients and outpatients, prisoners, etc.) on the validity and clinical scales. In 8 of the 15 comparisons, blacks got significantly higher *F* scores than whites. Although evidence of racial bias is most consistent in normal samples, there was no instance in any of the 25 studies (cf. Gynther, 1972; Gynther, Note 2) in which the scores of whites on this scale significantly exceeded the scores of blacks.¹

Should these results be taken as indicating that black subjects have more difficulties in completing the MMPI validly due to poor reading skills, inability to follow directions, confusion, or other sources of profile invalidity, or that black subjects are more rebellious, nonconforming, or emotionally disturbed than whites? That is, should the white-derived descriptors be applied to blacks? Only two studies have addressed this question. Gynther et al. (1973) found *no* correlates associated with $F \geq 26$ MMPIs produced by black psychiatric patients, although a meaningful cluster of correlates for this code type was established for white psychiatric patients. More recently, Hedlund (1977) showed that MMPI *F* scores of black psychiatric patients are positively associated with scores of items assessing "disorientation," "confusion," "delusions," and other behaviors suggesting psychosis. Hedlund (1977) pointed out that as far as Scale *D* and perhaps Scale *F* correlates were concerned, "most of the relationships were similar for blacks and whites but that smaller *N*s for the black samples precluded cross-validated statistical significances" (p. 744). Hedlund, however, con-

¹ Two additional black-white comparisons have been located. One used male misdemeanor offenders as subjects, and the other used male drug abusers. Blacks obtained nonsignificantly higher *F* scores than whites in the former study (McCreary & Padilla, 1977), but they obtained significantly lower *F* scores than whites in the other study (Penk & Robinowitz, 1974). This latter finding is unique among the 27 studies examined.

cluded that "very few of the significant white relationships were validated for blacks" (p. 744). In what appears to be a highly pertinent observation, Dahlstrom et al. (1972) stated that

an important part of interpreting test items seems to be an ability to share the common cultural framework of the derivational samples studied by the test authors. Many subjects lacking this essential experiential core will show their atypical attributions and self-labeling by endorsing *F*-scale items in an unusual way. (p. 117)

This study proposes to determine which of the 566 MMPI items were rarely (i.e., 10% or less) endorsed by a large sample of normal black adults. These items are then compared with *F*-scale items (and the 38 other rarely endorsed items mentioned earlier) that were derived from Minnesota whites using the same criterion. If there is a large amount of overlap between the two sets of items, support would be provided for the continued use (and general interpretation) of the MMPI *F* scale for blacks as well as for whites. If, on the other hand, the two *F* scales are quite different, it would be suggested that use of the original *F* scale with blacks leads to inaccurate and inappropriate representations of such subjects. These analyses were carried out as part of a large-scale investigation of the need for separate norms for ethnic minorities on the standard MMPI scales. This project is briefly described below, but a more complete report is in preparation (Lachar, Gynther, & Dahlstrom, Note 3).

Method

Subjects were obtained from Alabama, Michigan, and North Carolina to derive norms for blacks on the validity, clinical, and other frequently used scales of the MMPI (Lachar et al., Note 3). Major sources were church groups and social clubs, which were paid \$5 per participant. Black faculty and graduate students administered the inventory, usually to 5-10 persons at a time, and also asked subjects to supply face-sheet information on a form devised for that purpose. Confidentiality was assured by the use of code numbers rather than names. Total sample size was 882, 321 males and 561 females. Age varied from 18 to 65, with means 36.0 for males and 34.1 for females. Approximately 15% of the subjects had not completed high school, 27% were high school graduates, 26%

had some college, 18% had graduated from college, and 14% had some postgraduate work. Approximately two thirds of the subjects had attended segregated schools. Slightly over 50% of the sample was married, 30% were single, and the remainder were widowed, divorced, or separated. In terms of employment status, about 15% were unskilled, unemployed, or on welfare; about 21% were semi-skilled, approximately 13% were classified as skilled manual, nearly 18% were clerical workers, salespersons, or technicians; about 8% were administrative or minor professionals, 22% were managers or lesser professionals; and 3% were executives or major professionals. Mean annual family income for the males was \$14,756 and for the females was \$12,296. In response to questions concerning history of treatment for emotional (nervous) condition or time served in prison, 3.7% of the males responded affirmatively to the former question and 2.1% to the latter question. The comparable figures for females were 6.2% and .7%.

Comparison of sample demographic characteristics with contemporary census data (U.S. Bureau of the Census, 1976, 1977) suggests that the distribution of age, marital status, and percentage of unemployment is comparable to national estimates of blacks 18 years and older. Examination of other variables indicates that our sample had less frequently failed to complete high school (17% vs. 50%) and had attained more college experience (50% vs. 19%). A correlate of this greater education attainment in our sample was, in comparison to census data, less unskilled and more managerial and professional vocations, as well as a higher mean yearly family income. This substantially middle-class sample was the result of sample selection methodology that for the most part included socially active and community-oriented individuals. These biases in our sample were generally of the kind that would make our estimates of the presence of race-related differences at the item level conservative rather than extreme; that is, a more representative, less educated group of subjects would presumably respond in an even more discrepant manner.

Endorsement rates of all MMPI items were examined to locate those items endorsed, either positively or negatively, by 10% or less of the black sample. One could have analyzed male and female data separately or could have applied the 10% criterion to all 882 subjects regardless of sex. The former procedure was used to develop the MMPI *F* scale with the 111 males and 118 female protocols then available. If one looks at the endorsement data for the larger, revised Minnesota adult group (Dahlstrom et al., 1975, Appendix A), it will be noted that both males and females met the 10% or less criterion for 47 of the 64 *F* items. With regard to the remaining items, males exceeded the criterion on 12 items, females on 4 items, and males and females on 1 item. However, if the responses of both sexes are combined, the criterion was met for 60 of the 64 *F* items. Since the MMPI *F* scale

can best be characterized as consisting of items rarely endorsed in the scored direction by the *whole* sample (despite the procedure originally used), we felt that a black *F* scale to be comparable should also be determined by responses of the entire sample. However, if marked differences between the sexes on scale means (and standard deviations) were found, different raw-score to *T*-score conversions would be considered.

Comparisons were made between blacks and whites on the items that met the 10% or less criterion for blacks, items on the *F* scale that met the criterion, items on the *F* scale that did not meet the criterion, and items not on the *F* scale that met the criterion for the revised white adult group.

Results and Discussion

Examination of the endorsement rates of the black sample disclosed 33 items that met the 10% or less criterion. Twenty of these items are keyed true (10, 14, 23, 48, 49, 85, 104, 123, 151, 197, 210, 211, 227, 246, 291, 324, 339, 365, 393, 565), and 13 are keyed false (2, 75, 88, 90, 113, 177, 196, 220, 257, 258, 272, 276, 285). Although a number of statements that would be rated "obvious" for presence of psychopathology if answered true can still be found in this *F* scale for blacks (e.g., "I believe I am being followed," "Someone has been trying to poison me," and "Everything tastes the same."), it is noteworthy that a number of old standbys have dropped out (e.g., "My soul sometimes leaves my body," "I see things and animals around me that others do not see," and "I commonly hear voices without knowing where they come from.").

The scores of the black normative group on this scale can be summarized as follows: for males, $M = 2.97$, $SD = 4.02$, and range = 0-20; for females, $M = 2.72$, $SD = 3.52$, and range = 0-22. These values for males and females are sufficiently similar to make separate raw-score to *T*-score conversions unnecessary. A raw score of 3 can be considered equivalent to a *T* score of 50; a raw score of 7, equivalent to a *T* score of 60; and a raw score of 11, equivalent to a *T* score of 70. An additional cohort of subjects, black psychiatric patients evaluated at Lafayette Clinic, obtained the following scores: for 197 males (M age = 29.04; M education = 11.93), $M = 5.09$, $SD = 5.16$, and range =

0-33; for 280 females (M age = 31.17; M education = 12.31), $M = 5.92$, $SD = 4.95$, and range = 0-33. The distributions obtained by normal and psychiatric black subjects range more widely over this scale than do distributions of whites' scores over the standard *F* scale. It seems clear that some test-taking differences may still be reflected in the way that black subjects describe themselves, even on this selected set of items.

Relations among selected demographic variables, standard *F*-scale, and black *F*-scale scores were examined. Mean black *F*-scale *T* scores for various demographic groupings obtained a limited range of 47 to 56, whereas standard *F*-scale *T* scores ranged from 53 to 75. Analyses of variance revealed that scores on both scales were significantly related to age, education, and job classification of head of household. Standard *F*-scale *T* scores, for example, averaged less than 60 for those of our normal black subjects who were either 35 or older, had attained a college education, had managerial or professional status, or lived in a household headed by a managerial or professional person. If, on the other hand, the subject was 18-24 years of age, had less than 12 years of education, was unskilled or unemployed, or lived in a household with an unskilled or unemployed breadwinner, mean *F*-scale *T* scores ranged from 67 to 75.

Table 1 summarizes the endorsement rates of blacks and whites on the *F* items for blacks that were enumerated earlier in this section, the MMPI *F* items, and the 38 items that met the criterion for inclusion in the *F* scale but were not so classified by the test authors. Examination of these data shows that whites and blacks responded quite similarly to the 33 items that comprise the *F* scale for blacks. Although these figures suggest a high degree of overlap between responses to the *F* scale of blacks and responses to a substantial proportion of what appears to be the MMPI *F* scale, it should be noted that 6 of the 28 items do not appear on the original *F* scale. Comparison of the responses of blacks and whites to the standard 64-item scale shows considerable disagreement. Blacks exceeded the cutoff on 65.6%

Table 1
Number of Blacks' F Scale, MMPI F Scale, and Supplementary F Items Endorsed Below and Above the 10% Level by Minnesota White Normal Adults and Black Adult Subjects in the New Normative Sample

Group	F items					
	Blacks (<i>n</i> = 33)		MMPI (<i>n</i> = 64)		Supplementary (<i>n</i> = 38)	
	≤10%	>10%	≤10%	>10%	≤10%	>10%
Blacks	33	0	22	42	6	32
Whites*	28	4	60	4	38	0

Note. MMPI = Minnesota Multiphasic Personality Inventory.

*One of the 33 items (i.e., 565) cannot be assigned to either category, since the exact percentage of true responses is not available for the revised Minnesota white adult group.

of the items. Excluding the 4 items that did not meet the criterion for whites, there was agreement on 36.7% of the items. Performance on the supplementary *F* items was even more disparate; blacks and whites agreed, in the sense of both meeting the 10% or less criterion, on only 15.8% of the items. Although whites were in agreement with 85% of the items designated as *F* for blacks, blacks concurred with only 28 out of the 98 items that meet the 10% criterion for whites. Actual percentage differences between blacks and whites on the 70 items responded differently ranged up to 30. The mean percentage of difference for these items was 10.

There are many reasons for these differences. One explanation appears to be intrinsic to the original derivation of the MMPI *F* scale. The rarity of the deviant responses ranged from 0% (every white female responded true to Item 177, "My mother was a good woman") to 10% (and in four cases to above 10%, as we have indicated). One might ask if the endorsement rates of blacks agreed or disagreed with those of whites as a function of degree of rarity. Perhaps the responses of blacks to these items only differed from those that were "marginal" members of the category. To test this hypothesis, the *F* items were divided into two categories, one with a 6% or more endorsement rate (*n* = 34) and the other with a 5% or less endorsement rate (*n* = 30). Of the 22 items that blacks endorsed 10% or less, 18 fell in the 5% or less white endorsement category,

whereas 30 of the 42 items that blacks endorsed more than 10% fell in the 6% or more white endorsement category, $\chi^2(1) = 16.5$, $p < .01$. It would be simplistic (and not true) to say that blacks only disagreed with marginal *F* items, but the trend is definitely in that direction.

These findings can be viewed as what one might expect from a validity generalization study, that is, the typical reduction in number of significant items, descriptors, or correlates when original relationships are examined via a new sample. Would *any* new sample lead to similar results? Might not comparison of whites currently residing in Alabama, Michigan, and North Carolina with the vintage Minnesota whites lead to results similar to those found here? This is an empirical question, and we hope someone will collect the data to answer it. Our speculation, however, is that a present-day group of whites will not differ as much from the Minnesota norm group as our blacks did, because mean MMPI *F* scores of white normals have not changed much from the established norms over the years, whereas blacks' mean *F* scores have repeatedly been found to be at least as high and often substantially higher than the norms.

Since the degree of overlap between blacks' and whites' endorsement rates of MMPI *F*-scale items was only about 35%, it would appear that the newly derived *F* scale may be, especially for middle-class members of this minority group, a better measure of

deviant responding (and perhaps nonconformity?) than the standard scale. Whether blacks who obtain elevated scores on this new *F* scale are as moody, opinionated, rebellious, and emotionally disturbed as whites who obtain elevated scores on the standard *F* scale have been shown to be must await further research.

Reference Notes

1. Gough, H. G., McKee, M. G., & Yandell, R. J. *Adjective Check List analyses of a number of selected psychometric and assessment variables* (Tech. Rep. OERL-TM-55-10). Maxwell Air Force Base, Ala.: Officer Education Research Laboratory, 1955.
2. Gynther, M. D. *Ethnicity and personality: Recent findings*. Paper presented at the 12th annual MMPI Symposium, St. Petersburg, Florida, February 1977.
3. Lachar, D., Gynther, M. D., & Dahlstrom, W. G. *Personality inventory interpretation and the minority client: Development and application of MMPI black norms*. Book in preparation, 1977.

References

- Carson, R. C. Interpretive manual to the MMPI. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications*. New York: McGraw-Hill, 1969.
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook Vol. 1: Clinical interpretation*. Minneapolis: University of Minnesota Press, 1972.
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook Vol. 2: Research applications*. Minneapolis: University of Minnesota Press, 1975.
- Duckworth, J., & Duckworth, E. *MMPI interpretation manual for counselors and clinicians*. Muncie, Ind.: Accelerated Development, 1975.
- Graham, J. R. *The MMPI: A practical guide*. New York: Oxford University Press, 1977.

- Gynther, M. D. The clinical utility of "invalid" MMPI *F* scores. *Journal of Consulting Psychology*, 1961, 25, 540-542.
- Gynther, M. D. Crime and psychopathology. *Journal of Abnormal and Social Psychology*, 1962, 64, 378-380.
- Gynther, M. D. White norms and black MMPIs: A prescription for discrimination? *Psychological Bulletin*, 1972, 78, 386-402.
- Gynther, M. D., Altman, H., & Warbin, R. W. Interpretation of uninterpretable MMPI profiles. *Journal of Consulting and Clinical Psychology*, 1973, 40, 78-83.
- Gynther, M. D., & Shimkunas, A. M. More data on MMPI *F* > 16 scores. *Journal of Clinical Psychology*, 1965, 21, 275-277.
- Hathaway, S. R., & Briggs, P. F. Some normative data on new MMPI scales. *Journal of Clinical Psychology*, 1957, 13, 364-368.
- Hathaway, S. R., & McKinley, J. C. *The MMPI manual*. New York: Psychological Corporation, 1951.
- Hedlund, J. L. MMPI clinical scale correlates. *Journal of Consulting and Clinical Psychology*, 1977, 45, 739-750.
- Lachar, D. *The MMPI: Clinical assessment and automated interpretation*. Los Angeles: Western Psychological Services, 1974.
- McCreary, C., & Padilla, E. MMPI differences among black, Mexican-American, and white male offenders. *Journal of Clinical Psychology*, 1977, 33, 171-177.
- Meehl, P. E., & Hathaway, S. R. The *K* factor as a suppressor variable in the MMPI. *Journal of Applied Psychology*, 1946, 30, 525-564.
- Penk, W. E., & Robinowitz, R. MMPI differences of black and white drug abusers. *JSAS Catalog of Selected Documents in Psychology*, 1974, 4, 50. (Ms. No. 630)
- United States Bureau of the Census. *Statistical abstract of the United States*, 1976. Washington, D.C.: U.S. Government Printing Office, 1976.
- United States Bureau of the Census. Educational attainment in the United States: March 1977 and 1976. *Current population reports* (Series P-20, No. 314). Washington, D.C.: U.S. Government Printing Office, 1977.

Received November 28, 1977 ■

Parental Personality Factors in Child Abuse

John J. Spinetta
San Diego State University

In an attempt to demonstrate that abusing parents differ from nonabusing parents in personality variables, the Michigan Screening Profile of Parenting was administered to six groups of mothers: (a) adjudicated abusers, (b) spouses of adjudicated abusers, (c) mothers convicted of child neglect, (d) nonabusing mothers from a college student population, (e) nonabusing mothers from a middle socioeconomic level, and (f) nonabusing mothers from a lower socioeconomic level. Major differences occurred when comparison was made of one or more of the first three groups with one of the latter three groups. The groups differed significantly on six factor-analyzed cluster categories: (a) relationship to one's own parents, (b) tendency to becoming upset and angry, (c) tendency toward isolation and loneliness, (d) expectations of one's own children, (e) inability to separate parental and child feelings, and (f) fear of external threat and control. In all of the cases, the first three groups scored at levels of higher risk than did the latter three groups, whereas the abusers scored at the highest risk levels throughout. It is suggested that a therapist who helps a parent develop the ability to maintain equanimity under stress, by helping reduce deviations from the norm in characteristics related to abuse potential, is ultimately helping to reduce actual abusive behavior.

With the growing emphasis in the literature on the fact that the causes of child abuse are multiple and interactive, many therapists who deal with parental personality and attitudinal variables are made to feel as if they are engaging in a futile effort (D'Agostino, 1975; Smith, 1975). Although many new and exciting identification and treatment programs for child abuse abound throughout the country (National Center on Child Abuse and Neglect, 1975, 1976), very little encouragement has been given to the therapist who does not have easy access to the new interdisciplinary treatment programs and who, in many instances, remains the

sole therapeutic agent for a particular set of families (Steele, 1975). The problem is viewed as sufficiently complex that an individual therapist who deals solely with parental attitudes is often discouraged. It is the purpose of this study to demonstrate that parental personality and attitude are important factors in the etiology of child abuse. Such a demonstration can give hope to the therapist that efforts in dealing with the parental personality are aimed in a profitable direction and that he or she can be effective in reducing potential for abuse.

It is not my intent to suggest that factors of parental background or inadequacy are the sole determinants of child abuse. The fact is that the causes of child abuse are multiple and interactive; there is no single type of child abuser or a single causative factor as sufficient explanation of abuse (Spinetta & Rigler, 1972). Emphasis on parental personality is in no way meant to detract from these other factors. Rather, it is suggested that helping the parent to develop the ability to maintain equanimity

The author wishes to thank Richard Bourke, Esther Cardall, and Don King of the Department of Public Welfare, County of San Diego, and their social service staffs for their assistance in the administration of the questionnaires and Ruth Reinman and Richard Sprigle for their help in the coordination and analysis of the data.

Requests for reprints should be sent to John J. Spinetta, Department of Psychology, San Diego State University, San Diego, California 92182.

under stress is directly related to situational variables, and it can be of central value in the rehabilitative or preventive process.

It is in the broader context of situational variables that I ask the question, Why is it that the majority of parents do not abuse their children? Although in the socially and economically deprived segments of the population there is generally a higher degree of the kinds of stress factors found in abusing families, the great majority of deprived families do not abuse their children. Why is it that most deprived families do not engage in child abuse, though they are subject to the same economic and social stresses as those families who do abuse their children? Is there an actual difference between the types of stresses encountered by abusing parents and nonabusing parents within the same socioeconomic level (Gil, 1970, 1976), or is the difference in the parents' manner of approaching the stress situation (Kent, 1976; Smith, 1975; Spinetta & Rigler, 1972; Young, 1976)? I hold the latter position. When one takes into account the fact that some well-to-do and middle-class families also engage in child abuse, then one must look for the causes of child abuse beyond mere socioeconomic stress. The problem of etiology remains insoluble at the demographic level alone.

The present study is an attempt to demonstrate that however one might explain the particular circumstances that helped shape the parents' personality, abusing parents differ from nonabusing parents in attitudinal and personality variables.

Method

Instrument

In 1972, Schneider, Helfer, and Pollack disclosed efforts under way to design and validate a questionnaire with the goal of uncovering parents who have a potential to abuse their small children. They based their questions on their clinical experience, which suggested that parents who abuse their small children reported more severe physical punishment in their own childhood, more anxiety about dealing with their children's problems, more concern about being alone and isolated, more concern with criticism, and higher expectations for performance in their children than did nonabusers. After several

years of analysis and validation, they published first a 74-item and then a 50-item instrument, originally entitled *Survey on Bringing Up Children* (Schneider, Hoffmeister, & Helfer, 1976). The instrument has since been renamed the Michigan Screening Profile on Parenting (Helfer, Schneider, & Hoffmeister, 1977).

Although the questionnaire has not yet been sufficiently validated to be of use as a legally valid criterion in decisions regarding child placement or parental readiness to resume parenting functions, it has been shown to be capable of differentiating between attitudes regarding child rearing and regarding self-awareness and self-control functions in the parents.

With the permission of Helfer, I administered the questionnaire to several groups of parents, as discussed below, to see (a) whether abuse-potential cluster categories similar to those found by Helfer and his associates could be validated in a local sample and (b) whether scores based on the locally factor-analyzed categories could sort out abusing from nonabusing parents.

Subjects

As is typical of parents who come to the attention of public agencies (National Center on Child Abuse and Neglect, 1975), the parents referred to the participating agencies were from low socioeconomic levels. The use of such parents in the present study is not meant to suggest that abuse takes place only at low socioeconomic levels, because it does not (Spinetta & Rigler, 1972). Similarly, although more women than men have been found to abuse their children (Gelles, 1973; Gil, 1970; Smith, 1975), child abuse is not an act solely of the mother. However, the questionnaire was administered only to women to ensure nonconfounding by differences in child-rearing attitudes between men and women.

Subjects were chosen in the following manner: The participating agencies agreed to administer the questionnaire to all of the mothers currently under their jurisdiction as active cases. The questionnaire was administered to (a) adjudicated abusers, (b) spouses of adjudicated abusers, and (c) parents convicted of child neglect. The parents in these categories were chosen by the following criteria: (a) The child was under 5 years of age, and (b) court adjudication had been finalized, so that parents would not feel that their answers would affect the placement of their child or decisions regarding their own disposition. In this manner, workers were able to ensure that responses to the questionnaire were given as honestly as possible.

For purposes of comparison and contrast, the questionnaire was also administered to groups of parents who were nonabusers with children under 5 years of age. The following groups were tested: (d) nonabusing mothers from a college student population whose children were in a day-care center because one or both parents were in school, (e) nonabusing mothers from a middle socioeconomic level

whose children were in a preschool not because of necessity but through express parental wish, and (f) nonabusing mothers from a lower socioeconomic level with children in a preschool because the mother was working. Group f was chosen to match as closely as possible the educational, occupational, and socioeconomic status of Groups a, b, and c. Group d was chosen because it was similar to Groups a, b, and c in financial status but not in terms of education or potential occupation. Group e, different in terms of education, occupation, and financial status, and the most representative of the population as a whole, was chosen to test possible class differences in responding.

The samples consisted of the following numbers: (a) adjudicated abusers, 7; (b) spouses of abusers, 9; (c) parents convicted of neglect, 13; (d) nonabusing mothers from a college population, 15; (e) nonabusing mothers from a middle socioeconomic level, 15; and (f) nonabusing mothers from a lower socioeconomic level, 41.

The purpose of the study was explained in detail to the respective supervisors, the agency officials in Groups a-c, and the day-care administrators and teachers in Groups d-f. Because of the sensitive nature of the accusation of child abuse and neglect, and to prevent socially desirable responses, parents were not told specifically that the survey's ultimate purpose was to differentiate abuse potential. Rather, parents were asked if they wished to take part in a survey on attitudes in bringing up children, conducted by the university to learn how parents viewed child rearing. In accord with U.S. Department of Health, Education, and Welfare guidelines, parents were promised that the results would remain anonymous, and that any parent who wished would be given the overall results on completion of the study.

All of the parents who were approached in Groups d and e, without exception, filled out the survey as requested. Of the parents approached in Group f, all but three (93%) filled out the survey. The parents in Groups a-c were approached by assigned workers who had established rapport with them and were told that this survey would not only aid the university but that it might be of therapeutic aid to the specific worker in each case. Each worker was asked to screen out those parents who would be unduly threatened by the questionnaire, those who might be tempted to answer with socially desirable responses, and those whose cases were still pending court completion. The workers did not receive any refusals from the selected cases. The final small sample thus represents responses from parents who were motivated to fill out the questionnaires as honestly as possible. Comments from each worker on each case attested to the honest efforts of the parents who made up the final samples in Groups a-c. It is my belief that the final sample represents the cases most amenable to treatment. There is no reason to suspect that the sample represents the most severe of the abusers. On the contrary, workers' case records show that the final sample is

on the conservative side of the abuse-potential continuum in the agencies' overall abuser population. Thus, any differences that appear between the abuser and nonabuser groups would appear at least equally as strong in the general abuser population of the agencies in question. With the questionnaire aimed at being of eventual use as an aid to the therapist in sorting out areas of weakness, honest cooperation of the parents was deemed essential. In addition, honest cooperation in each of the six groups minimized confounding that would appear if the groups differed in willingness to participate.

Results

A varimax rotated factor analysis of the responses to the questionnaires was conducted by the experimenter. The six clusters of variables closely resemble the high-abuse potential categories of Helfer et al. (1977). The six resultant clusters of the present analysis are (a) relationship to one's own parents, (b) tendency to becoming upset and angry, (c) tendency toward isolation and loneliness, (d) expectations of one's own children, (e) inability to separate parental and child feelings, and (f) fear of external threat and control.

With these six factor-analyzed cluster categories as a basis, a six-column scoring form was devised, with direction of scoring set so that the higher score on each cluster represented abuse potential. Total raw scores for each subject were determined for each of the six cluster categories.

A 1×6 analysis of variance was performed for the six groups for each of the six abuse-potential categories. Table 1 gives the means and standard deviations for scores in each of the abuse-potential categories for each subject group. Table 2 gives the results of the analysis of variance for each of the six categories.

Scores on each of the six abuse-potential categories showed that significant differences existed among the six groups ($df = 5, 90$ in all cases). The resultant F on the first abuse-potential category, relationship to one's own parents, was 4.55, significant at the .001 level. The resultant F of 6.70 on the second abuse potential category, tendency to becoming upset and angry, was significant at the .001 level. The resultant F on the third category, tendency toward isolation and lone-

Table 1

Means and Standard Deviations in Each Abuse Potential Category

Cluster	1: Abusers		2: Spouses		3: Neglect		4: College		5: Middle		6: Lower	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
1 (Parents)	57.4	14.7	48.7	9.4	53.3	10.3	44.9	11.6	37.7	10.2	44.3	10.3
2 (Control)	25.4	9.6	22.2	7.6	22.8	7.3	17.7	4.0	14.1	3.8	16.7	4.7
3 (Affiliation)	31.9	8.2	26.9	5.3	25.9	4.0	22.5	4.2	19.9	3.8	22.5	4.8
4 (Expectations)	39.1	18.6	37.3	11.7	34.3	10.7	28.7	7.8	22.3	6.3	30.0	8.8
5 (Symbiosis)	17.3	5.1	16.2	2.3	19.2	2.9	14.9	2.1	14.5	2.7	16.2	3.3
6 (Threat)	61.3	16.5	52.6	12.9	57.4	10.8	40.7	8.7	29.3	5.7	43.9	10.5

liness, was 7.53, significant at the .001 level. The resultant *F* on the fourth category, expectations of one's own children, was 4.20, significant at the .001 level. The resultant *F* on the fifth category, inability to separate parental and child feelings, was 3.79, significant at the .01 level. The resultant *F* of 13.92 on the sixth abuse-potential category, fear of external threat and control, was significant at the .001 level.

A posteriori tests using the Scheffé method were conducted for each of the abuse-potential clusters. Significant differences were found as follows: Group a (abusers) sig-

nificantly differed from Group e (middle-class nonabusers) in Abuse-Potential Clusters 1, 2, 3, 4, and 6. Group a significantly differed from Groups d and f in Abuse-Potential Clusters 2, 3, and 6.

Group b (spouses of abusers) significantly differed from Group e in Abuse-Potential Clusters 2, 3, 4, and 6.

Group c (neglecters) significantly differed from Group e in Abuse-Potential Clusters 1, 2, 5, and 6. Group c significantly differed from groups d and f in Abuse-Potential Clusters 2, 5, and 6.

The Scheffé a posteriori test showed that the major differences in each of the six abuse-potential categories occurred when comparison was made of one or more of the first three groups (abusers, abusers' spouses, and neglecters) with one of the latter three groups (nonabusers). The greatest differences occurred when each of the first three groups was compared to the fifth group (middle-class nonabusers). In each of the abuse-potential categories, Group e scored at the lowest level. Group d (college student nonabusers) and Group f (lower socioeconomic level nonabusers) were the next lowest in abuse potential, scoring almost identically throughout. Although the fifth group scored lowest on all of the categories, the other two nonabuser groups scored at a level not significantly higher. In contrast, the abusers scored at the highest risk level in all but one of the abuse-potential categories.

Table 2

Analysis of Variance

Cluster	<i>MS</i>	<i>F</i>
1		
Between	527.5	4.55**
Within	116.1	
2		
Between	213.5	6.79**
Within	31.5	
3		
Between	177.3	7.53**
Within	23.5	
4		
Between	409.3	4.20**
Within	97.5	
5		
Between	35.9	3.79*
Within	9.5	
6		
Between	1,546.3	13.92**
Within	11.1	

Note. *df* = 5, 90.

* *p* < .01.

** *p* < .001.

Discussion

The Michigan Screening Profile on Parenting was able to differentiate between abusing

and nonabusing mothers on personality and attitudinal variables. The empirically derived set of abuse-potential categories proved useful in significantly differentiating between abusing and nonabusing mothers within the same socioeconomic level in three areas: the tendency to becoming upset and angry, feelings of isolation and loneliness, and the fear of external threat and control. The abusing mothers differed significantly from nonabusing mothers in a middle socioeconomic level in the same categories; in their relationship to their own parents, both past and present; in having higher than normal expectations for their young children's performance; and in failing to separate their own feelings from those of their children. Although not at a significant level, abusing mothers differed from nonabusing mothers in the same socioeconomic level in the latter categories as well. Neglecting parents and spouses of abusers were also shown to be weak in the six abuse-potential categories.

Personality and attitudinal factors do make a difference. Abusing mothers differ from nonabusing mothers in areas of attitude and personality that have been clinically related to potential for abuse (Colman, 1975; Corey, Miller, & Widlack, 1975; Kent, 1976; Paulson et al., 1974; Smith, 1975; Spinetta & Rigler, 1972; Steele, 1975; Tracy & Clark, 1974; Walters, 1975). The fact that neglecting mothers and spouses of abusers also scored high on the abuse-potential categories demonstrates the power of the test in pointing to weaknesses in parental personality and attitudes that can affect the parenting role itself, regardless of whether the result is actual physical abuse, neglect of the child, or passively allowing one's spouse to abuse the child. Intervention and direction is called for in each case.

As stated above, there is no suggestion made that factors of parental inadequacy and personality weakness are the sole determinants of child abuse. Certainly, those involved in the care of the abusing parent must continue to relieve the family as much as possible of overwhelming situational stresses. However, personality does play a role. The therapist who helps the parent de-

velop the ability to maintain equanimity under stress, can be of immense aid in the rehabilitative or preventive effort.

One must caution that the questionnaire cannot be used as a legally valid criterion for sorting out abusing from nonabusing parents, since false positives have been shown on occasion (Schneider et al., 1976) and since false negatives can appear with those parents who refuse to answer the questions honestly. It is possible to fake answers by giving socially desirable responses. However, for those parents in a therapeutic situation who respond to the questionnaire with an honest desire to be helped, the responses can help point to weaknesses in areas that have been clinically shown to relate to potential for abuse. A therapist who directs interventional and preventive efforts toward the amelioration of parental attitudes, both attitudes toward the self and toward the child, is not, as Alby (1975) suggested, misdirecting energies, but is rather helping reduce deviations from the norm in characteristics related to abuse potential and, hopefully, is ultimately helping reduce the actual abusive behavior.

References

- Alby, K. T. Preventing child abuse. *American Psychologist*, 1975, 30, 921-928.
- Colman, W. Occupational therapy and child abuse. *American Journal of Occupational Therapy*, 1975, 29, 412-417.
- Corey, E. J., Miller, C. L., & Widlack, F. W. Factors contributing to child abuse. *Nursing Research*, 1975, 24, 293-295.
- D'Agostino, P. Strains and stresses in protective services. In N. B. Ebeling & D. A. Hill (Eds.), *Child abuse: Intervention and treatment*. Acton, Mass.: Publishing Sciences Group, 1975.
- Gelles, R. J. Child abuse as psychopathology: A sociological critique and reformulation. *American Journal of Orthopsychiatry*, 1973, 43, 612-621.
- Gil, D. G. *Violence against children: Physical abuse in the United States*. Cambridge, Mass.: Harvard University Press, 1970.
- Gil, D. G. Primary prevention of child abuse: A philosophical and political issue. *Pediatric Psychology*, 1976, 1(2), 54-57.
- Helper, R. E., Schneider, C., & Hoffmeister, J. K. *Manual for use of the Michigan Screening Profile of Parenting*. East Lansing: Michigan State University Press, 1977.
- Kent, J. T. A follow-up study of abused children. *Pediatric Psychology*, 1976, 1(2), 24-27.

- National Center on Child Abuse and Neglect. *Child abuse and neglect: The problem and its management* (3 vols) (DHEW Publ. No. OHD 75-30073). Washington, D.C.: U.S. Govt. Printing Office, 1975.
- National Center on Child Abuse and Neglect. *Federally funded child abuse and neglect projects 1975* (DHEW publ. No. OHD 76-30076). Washington, D.C.: U.S. Government Printing Office, 1976.
- Paulson, M. J., et al. Parents of the battered child: A multidisciplinary group therapy approach to life-threatening behavior. *Life-Threatening Behavior*, 1974, 4, 18-31.
- Schneider, C., Helfer, R. E., & Pollack, C. The predictive questionnaire: Preliminary report. In C. H. Kempe & R. E. Helfer (Eds.), *Helping the battered child and his family*. Philadelphia: Lippincott, 1972.
- Schneider, C., Hoffmeister, J. K., & Helfer, R. E. A predictive screening questionnaire for potential problems in mother-child interaction. In R. E. Helfer & C. H. Kempe (Eds.), *Child abuse and neglect: The family and the community*. Cambridge, Mass.: Ballinger, 1976.
- Smith, S. M. *The battered child syndrome*. London: Butterworth, 1975.
- Spinetta, J. J., & Rigler, D. The child-abusing parent: A psychological review. *Psychological Bulletin*, 1972, 77, 296-304.
- Steele, B. F. *Child abuse and neglect: The problem and its management* (Vol. 3, chap. 3, pp. 65-114) (DHEW Publ. No. OHD 75-30073). Washington, D.C.: U.S. Govt. Printing Office, 1975.
- Tracy, J. J., & Clark, E. J. Treatment for child abusers. *Social Work*, 1974, 19, 338-342.
- Walters, D. R. *Physical and sexual abuse of children: Causes and treatment*. Bloomington: Indiana University Press, 1975.
- Young, M. Multiple correlates of abuse: A systems approach to the etiology of child abuse. *Pediatric Psychology*, 1976, 1(2), 57-61.

Received November 29, 1977 ■

Effects of Intestinal Bypass Surgery on Body Concept

Peter M. Silberfarb, Patricia J. Phelps,
Peter Hauri, and Charles Solow
Dartmouth Medical School

Stability of body concept, as reflected in the Draw-a-Person Test, was demonstrated in 14 patients undergoing bypass surgery for severe obesity. Three Articulation-of-Body Concept scores for each patient were obtained: at the preoperative level, within 1 year subsequent to surgery, and at least 2 years after the operation. During the first year after the operation, body concept was significantly poorer than at the preoperative level. However, patients recovered from their postoperative disruption of body concept within 2 years after the operation and then reached at least a preoperative level. Thus, the shock of the bypass operation seemed to cause a temporary disruption in this otherwise extremely stable personality dimension.

Field dependency has become an important personality dimension when attempting to explain individual patterns and styles of behavior (Witkin, Dyk, Fatereson, Goodenough, & Karp, 1962; Witkin, Goodenough, & Karp, 1967). Field-dependent individuals, when presented with an organized Gestalt, appear to have difficulty separating out parts of that whole, whereas field-independent persons are more able to perceive the discrete parts of a complex pattern as separate from each other (Witkin et al., 1962). The degree of field dependency within a certain individual appears to manifest itself in drawings of the human figure. Specifically, field-dependent persons show less articulation of body parts in their drawings than do field-independent individuals.

An Articulation-of-Body Concept (ABC) scale for children has been developed and validated by Marlen's (Fatereson & Witkin, 1970;

Witkin et al., 1962). This scale basically reflects three characteristics: (a) the extent of identity and sex differentiation in the figure drawings, (b) the amount of detail in the drawings, and (c) the form level of the figures (i.e., general body shape and integration on features). Low scores indicate relatively primitive drawings, and higher scores are given for more sophistication. As predicted, children who score higher on the Marlen's scale generally are found to be more field independent on other measures as well, whereas children who draw relatively unarticulated figures are more field dependent (Witkin et al., 1962). Studies containing adults indicate similar relationships (Reitman & Cleveland, 1964; Witkin et al., 1962). Occasionally, the 5-point Marlen's ABC scale is extended to a 9-point scale in adult studies, and this 9-point version was used in the current evaluation.

An individual's level of field dependency appears to be stable. It is not changed by such interventions as sensory isolation (Reitman & Cleveland, 1964), electroconvulsive shock (Pollack, Kahn, Karp, & Fink, Note 1), and drugs, including chlorpromazine, imipramine, and alcohol (Karp, Witkin, & Goodenough, 1965; Pollack et al., Note 1). However, the question still remains as to how much the stability of a person's score on

This study was supported in part by Grant 1R03MH17294-1 from the National Institute of Mental Health.

We wish to thank Marlen's and Witkin for help in preparation of this manuscript and Swift for initial work in collection of the clinical data.

Requests for reprints should be sent to Peter M. Silberfarb, Department of Psychiatry, Dartmouth Medical School, Dartmouth College, Hanover, New Hampshire 03755.

the Marlens scale is dependent on the stability of physical body over time. Although some of the manipulations studied so far have threatened body appearance, none have actually changed an individual's body appearance in any dramatic way.

An ideal opportunity to study the stability of body articulation scores presented itself when drawings were collected from patients undergoing jejunoileostomy, an operation that drastically alters body appearance. Jejunoileostomy (intestinal bypass surgery for massive obesity) leads to a relatively consistent, substantial, and lasting weight reduction. Weight stabilizes after 1 or 2 years, usually at somewhat above ideal levels. The procedure, however, has distressing side effects and complications that are sufficiently common to require caution and restraint in the use of this therapeutic approach.

The specific question investigated in this study is whether the drastic and lasting changes in a person's body appearance secondary to jejunoileostomy lead to temporary or chronic changes in body concept as measured by the Marlens scale. This question seemed relevant, because field dependency (indirectly measured by the Marlens scale) has assumed a fair amount of importance in psychological testing and personality theory.

Method

Subjects

Twenty-nine patients were tested and interviewed before and after jejunoileostomy and at follow-up, at least 2 years later, as part of an investigation concerning the psychological sequelae of jejunoileostomy (Solow, Silberfarb, & Swift, 1974). Of these 29 patients, three scorable figure drawings (at preoperative, postoperative, and follow-up levels) were collected only from 14 patients. In 10 cases, the interviewer did not ask for figure drawings on all three occasions, because they had not been part of the test battery as originally planned. Five other patients drew stick figures on at least one of the three occasions. Such stick figures cannot be rated on the Marlens ABC scale.

Of the 14 patients who produced sets of three ratable figures, 10 were women and 4 were men. The ages ranged from 21 to 50 ($M = 36$ years). Preoperative weights ranged from 102 kg to 223 kg ($M = 151$ kg), representing from 39 to 160 kg over desirable weight (M overweight = 84 kg). Mean weight loss at follow-up was 60 kg, with a range of 32 kg to 117 kg. Only 2 of the patients had be-

come obese in adulthood; the others had been overweight since before the age of 16. All patients had been severely obese for years and had tried many diets without long-lasting success.

Eight of the patients in this study had been referred for jejunoileostomy by their physicians, and 6 were self-referred, although motivation was mixed in most cases. Five sought surgery primarily because of somatic concerns, whereas 9 were motivated primarily by psychosocial concerns such as the need to improve appearance, marriage, and so forth.

Nine patients were considered reasonably well adjusted before the bypass operation, and five were seen as distinctly impaired psychologically. Among the latter, four were diagnosed neurotic, and one was believed to have a personality disorder. Deficient self-esteem, marked self-consciousness about appearance, and vocational impairment characterized at least two thirds of the total group and appear to be related to the restriction of physical and social activity caused by their massive obesity.

Procedure

Each patient was interviewed three times by a trained psychiatrist, once at the preoperative level [within 1 week ($M = 4$ days) before surgery], once at the postoperative level [from 161 to 370 days ($M = 235$ days) after surgery], and once at follow-up (at least 2 years after the bypass operation). During each of these interviews, the patient was asked to draw a person. Interviewers did not know the purpose of this request and were unfamiliar with the theory behind the Marlens ABC scale.

For scoring, each drawing was randomly assigned a code number. All other identifying marks were removed from each drawing except for identifying the patient as either male or female. Marlens, the developer of the 9-point ABC scale, then rated the drawings in a random order. Marlens knew nothing about the purpose of this experiment except that all drawings came from adults.

Statistical Analysis

Ratings were evaluated by a two-way analysis of variance using a mixed repeated measures design. [The 14 patients were interpreted as a random variable, and the three conditions (preoperative, postoperative, and follow-up levels), as a fixed variable.] When the analysis of variance showed significance, differences among the three conditions were further evaluated by matched t tests. Finally, Pearson correlation coefficients were computed among the three conditions (preoperative, postoperative, and follow-up levels) and were adjusted for multiple comparisons by Sheffé's method.

Results

Table 1 indicates the mean ABC ratings: They were similar at preoperative and fol-

Table 1
Mean Articulation of Body Concept Ratings

Time of drawing	M ABC rating	SD
Preoperative ($M = 4$ days)	6.29	2.70
Postoperative ($M = 235$ days)	5.43	2.34
Follow-up ($M = 32$ months)	6.43	2.50

Note. ABC = Articulation-of-Body-Concept scale.

low-up levels but were somewhat lower (less well articulated) at the postoperative period. (For overall results of other interview data, see Solow et al., 1974).

Tables 2 and 3 (correlational analysis) document the consistency of ABC ratings over time. Most of the variance in Table 2 was due to variability from one patient to the other ($F = 18.40$, $p < .001$). Similarly, correlations from preoperative to postoperative and follow-up levels were highly significant, suggesting high test-retest reliability of the ABC scale over a fairly extended time period.

Within this overall stability of the ratings, there was a slight but nevertheless significant decrease of ABC scores postoperatively. This is indicated by the significant "time of drawing" effect in Table 2 ($F = 3.19$, $p < .05$) and by the t -test analyses in Table 3: Scores significantly decreased from preoperative to postoperative levels but increased again at follow-up.

Although the postoperative drawings were scored significantly lower than either the preoperative or the follow-up drawings, weight loss (either in percentages of kilo-

Table 2
Analysis of Variance for ABC Ratings

Source	SS	df	MS	F
Patients	221.62	13	17.05	18.40**
Time of drawing	5.90	2	2.95	3.19*
Residual	24.10	26	.93	—

Note. ABC = Articulation-of-Body-Concept scale.

* $p < .05$.

** $p < .001$.

grams) was not significantly correlated with the deterioration in ABC scores from preoperative to postoperative tests.

Discussion

The most important result of this study is the remarkable consistency in the patients' scores on the ABC scale from the preoperative level to follow-up, even though these patients underwent marked changes in their own body appearance. (Mean weight loss was 60 kg!) This consistency in ABC scores in the face of marked body alteration supports the hypothesis that the degree of articulation of body concept in adults is *not* directly related to the patient's physical appearance. In addition, this finding is made more interesting by the fact that the massive weight reduction undergone by these patients must have also affected their postural experience.

One rationale for using human figure drawings in psychological testing is that they reflect the patient's degree of field dependence. The present study suggests that in

Table 3
Comparisons Among the Three Sets of ABC Ratings

Sets compared	Time between compared sets	t	r
Preoperative vs. postoperative	$M = 239$ days	3.39***	.84****
Preoperative vs. follow-up	at least 2 years	.33	.81****
Postoperative vs. follow-up	at least 1 year	2.37*	.72**

Note. ABC = Articulation-of-Body-Concept scale.

* $p < .04$.

** $p < .01$.

*** $p < .005$.

**** $p < .001$.

adults, this concept is relatively fixed, because marked alteration in the physical shape and size of the body did not change the ABC scores at follow-up. If body concept were a more fluid entity, one might expect figure drawings to change in proportion to the amount of body change incurred after the bypass surgery. There was no correlation, however, between the change in ABC scores and total weight lost.

A second finding in this study lies in the fact that the jejunoileostomy did, temporarily, shift the patients toward less sophisticated figure drawings, possibly indicating a temporary shift toward more field dependence after the operation. It seems unlikely that reduced motivation to draw the figure could have played some role in causing less articulated drawings after surgery. High test-retest reliability commonly found for ABC scores suggest that body articulation is only minimally influenced by transient factors such as motivation (Faterson & Witkin, 1970). Also, it would be difficult to explain how reduced motivation could occur so consistently across subjects. Rather, it seems that the postoperative period of intestinal bypass surgery is a time of major readjustment, and one might speculate that this drastic alteration of one's body may have disrupted the body image temporarily.

In summary, jejunoileostomy in massively obese patients seemed to cause a temporary decrease but no permanent change in the articulation of body concept. Thus, field dependence seems to remain stable not only in relation to the effects of sensory isolation, drugs, and electroconvulsive therapy (as pre-

viously demonstrated) but to dramatic changes in actual body configuration as well. This stability is in sharp contrast to the changes noted in other psychological variables following jejunoileostomy surgery such as self-esteem and self-consciousness (Solow et al., 1974). This lends further support to the concept of field dependence as an enduring personality trait.

Reference Note

1. Pollack, M., Kahn, R. L., Karp, E., & Fink, M. *Individual differences in the perception of the upright in hospitalized psychiatric patients*. Paper presented at the meeting of the Eastern Psychological Association, New York, April 1960.

References

- Faterson, H. F., & Witkin, H. A. Longitudinal study of development of the body concept. *Developmental Psychology*, 1970, 3, 429-438.
- Karp, S. A., Witkin, H. A., & Goodenough, D. R. Alcoholism and psychological differentiation: The effect of alcohol on field dependence. *Journal of Abnormal Psychology*, 1965, 70, 4, 262-265.
- Reitman, E. E., & Cleveland, S. E. Changes in body image following sensory deprivation in schizophrenic and control groups. *Journal of Abnormal and Social Psychology*, 1964, 68, 168-176.
- Solow, C., Silberfarb, P. M., & Swift, K. Psychosocial effects of intestinal bypass surgery for severe obesity. *New England Journal of Medicine*, 1974, 290, 300-304.
- Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R., & Karp, S. A. *Psychological differentiation*. New York: Wiley, 1962.
- Witkin, H. A., Goodenough, D. R., & Karp, S. A. Stability of cognitive style from childhood to young adulthood. *Journal of Personality and Social Psychology*, 1967, 7, 291-300.

Received December 7, 1977 ■

Effects of Sex, Social Desirability, and Birth Order on the Defense Mechanisms Inventory

Gary E. Dudley
Georgia Southern College

This study investigated the effect of sex difference, social desirability instructions, and the birth order of respondents on the Defense Mechanisms Inventory (DMI). Using 30 male and 30 female undergraduates, half of the subjects in each group were given regular instructions, and the other half were instructed to respond so as to present a favorable impression. It was hypothesized that a sex difference would be found on Turning-Against-Others (TAO) and Turning-Against-Self (TAS) and that social desirability instructions would result in significant differences for TAO, Projection (PRO), Principalization (PRN), and Reversal (REV). It was further hypothesized that firstborns would report less TAO than later borns. In contrast to previously published reports on the DMI, a sex difference was found on PRO only. Further, social desirability effects were found on TAO, PRO, PRN, and REV. Thus, an interpretive caution is in order regarding the use of the DMI, but its potential clinical utility suggests that further research is warranted.

Although interest in and use of the notion of "defense mechanisms" is firmly ingrained in clinical practice, research on the cognitive and behavioral operations involved is remarkably absent in academic psychology. In a recent publication, Gleser and Ihlevich (1969) reported on the Defense Mechanisms Inventory (DMI). The DMI represents the first attempt to derive a comprehensive, objective, and behaviorally stated test of the traditionally defined mechanisms of defense. The test provides subjects with a checklist of possible situations, and the subjects are asked to indicate what would be their most likely and their least likely reactions. Scores on five defense clusters are derived from the responders' rankings: Turning-Against-Others (TAO), Projection (PRO), Principalization (PRN), Turning-Against-Self (TAS), and Reversal (REV). (See Gleser & Ihlevich, 1969, for a detailed description of each defense cluster.)

The DMI has found a useful role in the medical setting inasmuch as practitioners are generally interested in making a statement about the structure and function of psychological defenses in evaluating surgical and other medical patients. Two recently reported studies lend some support to the construct validity aspect of the DMI in this area. Gur and Gur (1975) reported that persons who scored high on REV (repressive/denial defenses) had significantly more psychosomatic complaints than persons who used affect-expressive defenses of TAO or PRO. Klein, Gonen, and Smith (1975) reported that high TAS and REV scores were consistent with the psychogenic diagnosis of a patient with painful ecchymosis following surgery for a herniated lumbar disc. Similarly, Scholz (1973) reported on 35 suicide attempters paired with 35 nonsuicidal neuropsychiatric patients. As hypothesized, suicide attempters displayed significant differences on the TAS dimension.

Psychometric data on the DMI suggest adequate reliability, although one study found a sex difference on three of the five defense scales, with males scoring higher on

This research was supported by Grant 0377 from Georgia Southern College.

Requests for reprints should be sent to Gary E. Dudley, Department of Psychology, Georgia Southern College, Statesboro, Georgia 30458.

TAO and PRO and females scoring higher on TAS (Weissman, Ritter, & Gordon, 1971). Gleser and Ihlevich (1969) originally reported a sex difference for TAS only. Finally, in another construct validity study, Gleser and Sacks (1973) reported that the DMI adequately predicted actual behavior in a conflict situation for males but not for females. One possible explanation for this finding is the influence of a social desirability factor on the females' reactions to the experimental manipulation. In fact, the social desirability aspects of the DMI have not been evaluated. Such an evaluation may help clarify the potential clinical utility of the DMI. Further, since published results are inconsistent with regard to sex differences, and since it is possible, for example, that the effects of sex and social desirability may have been influential in the Gleser and Sacks study, the DMI was evaluated to determine the maximum extent to which it may be influenced by social desirability in the context of sex differences. It was hypothesized that sex differences would be found for TAO (males higher) and TAS (females higher) and that main effects would obtain for social desirability instructions on TAO, PRO, PRN, and REV. Finally, firstborns were expected to use TAO less than later borns. This hypothesis was based on the findings of Warren (1966) that firstborns tend to be more concerned with social desirability than later borns and the view that TAO would be seen as a less socially desirable mode of conflict resolution. It also served as a more subtle indicator of the social desirability features of the DMI.

Method

Subjects

The study contained 60 undergraduate students from an introductory psychology class. There were an equal number of males and females ranging in age from 18 to 23 years ($M = 19.1$).

Procedure

All students completed the DMI and provided information about birth order and age. One group was given regular instructions as provided with the

DMI. The second group was given social desirability instructions. Each group consisted of 15 males and 15 females. The social desirability instructions consisted of a modification of Paragraphs 2 and 4 in the regular instructions. These paragraphs were modified to read as follows:

What we want you to do is to select the *one* answer of the five which you think is most *socially appropriate*. That is, select the answer which is most likely to create a favorable impression in the social situation, and fill in the box labeled "T" by the number corresponding to that answer on the attached answer sheet. Then select the *one* answer which you think is least favorable or least socially appropriate and fill in the box by that number labeled "F." Remember: You are *not* to answer in the way that you would necessarily react but in the way a person would react who is trying to create a favorable social image.

There are no right or wrong answers here; the only thing that should guide your selections is your knowledge about how to create a favorable image of yourself. Allow your mind to imagine for a moment that the event described in the story is really happening to you, even though you may never have experienced such an event. Remember, we are *not* asking what *your* behavior and responses would be but rather your opinion of what the likely behavior would be of someone trying to present themselves favorably in our society.

The data were analyzed by multivariate analysis of variance with sex and instructions as independent variables. A *t* statistic was derived to test the hypothesis of a birth-order difference on TAO.

Results

Multivariate analysis of variance on the five dependent measures resulted in a significant overall main effect for instructions, approximate $F(5, 52) = 12.20$, $p < .01$ (Dixon, 1973). The interaction was not significant. Subsequently, each dependent measure was evaluated in a univariate analysis of variance that produced significant main effects for sex on PRO, $F(1, 56) = 5.79$, $p < .05$, and for instructions on TAO, PRO, PRN, and REV, $F(1, 56) = 32.28$, $p < .001$, $F(1, 56) = 16.63$, $p < .001$, $F(1, 56) = 12.59$, $p < .001$, and $F(1, 56) = 52.28$, $p < .001$, respectively. Even though no interactions were significant, there was a tendency for males to decrease TAS under social desirability instructions, $t(58) = 1.4$, $p < .20$,

whereas females showed no change in TAS in the social desirability condition. Finally, firstborns scored significantly lower on TAO than later borns ($M_s = 32.1$ and 39.0 , $SD_s = 7.96$ and 7.01 , respectively) as predicted, $t(58) = 2.33$, $p < .05$, see Table 1).

Discussion

Based on the results of this study, it is clear that the question of sex differences in defensive style is unresolved. In contrast to two earlier studies (Gleser & Ihlevich, 1969; Weissman et al., 1971), the present study failed to demonstrate a sex difference for TAS. Further, the finding that males scored higher than females on PRO is consistent with Weissman et al., but their additional sex difference on TAO did not occur in this study. Some of these discrepancies may be a function of regional differences in the sample. Previous studies were done in Northern cities, whereas subjects in this study were primarily from the rural South. Specifically, the influence of the Southern Baptist religious orientation (40% of the sample) may have militated against the selection of TAO defenses in both sexes. The fact that PRO, related to TAO in terms of the direction of affective expression, was used more by males than females suggests that although males did not report a direct response against the frustrating object or person, they were willing to attribute negative characteristics to frustrating persons. This represents a more subtle form of aggression and externalization of negative affective experience. It should be noted that the present male sample had a mean TAO score of 33.6 compared to the mean TAO score of 40.9 reported for a college sample by Gleser and Ihlevich (1969). The absence of a sex difference for TAS in the present sample is inconsistent with all previously published reports. Higher intro-punitiveness in males may also be a function of the sample, however, and the effects of socioeconomic level, religious orientation, and demographic variables should be explored. Alternatively, one of the most serious shortcomings of the DMI in its present form is that it is an ipsative measure. Thus, given the data from this study, it is likely that one

Table 1
Means and Standard Deviations for DMI Subscales With Regular and Social Desirability Instructions

Instructions	Males		Females	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
DMI				
TAO	33.6	12.4	36.6	8.1
PRO	39.5	5.1	37.3	5.0
PRN	46.5	5.6	44.8	5.6
TAS	40.1	9.9	40.5	6.8
REV	40.7	7.9	40.9	8.5
Social desirability				
TAO	21.5	12.4	17.9	8.3
PRO	35.1	6.0	30.9	4.2
PRN	50.5	6.7	51.5	5.5
TAS	36.3	8.5	39.7	5.8
REV	56.6	10.5	60.0	10.3

Note. DMI = Defense Mechanisms Inventory; TAO = Turning-Against-Others; PRO = Projection; PRN = Principalization; TAS = Turning-Against-Self; REV = Reversal.

or the other of the discrepant findings (low TAO in males or high TAS in males) is related to the ipsative property of the instrument. (For a comprehensive treatment of this psychometric issue, the reader is referred to Block, 1957; Broverman, 1962; Cattell, 1944.)

The findings related to social desirability have important implications for the clinical utility of this instrument. The present sample readily identified the TAO and PRO items as less desirable than the PRN and REV items. Thus, if there is any motivation to present a favorable picture of oneself, it seems clear that the instrument in its present form does not provide adequate safeguards. A more subtle indicator of the social desirability effect is provided by the finding that firstborns, generally considered to be more inclined to respond in socially desirable ways (Warren, 1966), in fact scored lower on TAO (a defensive style that was avoided by persons given social desirability instructions) than did later borns.

In summary, the DMI is a promising instrument in research on defensive and coping processes in personality research. It seems almost certain that it will find increasing util-

ity in psychosomatic medicine (Gur & Gur, 1975; Klein et al., 1975). The present study provides data that suggest the need for caution, however, in using the instrument clinically because of apparent high susceptibility to the influence of social desirability response biases. Finally, the ipsative nature of the instrument in its present form has been noted, and the psychometric problems posed in this regard should be explored in subsequent research.

References

- Block, J. A. A comparison between ipsative and normative ratings of personality. *Journal of Abnormal and Social Psychology*, 1957, 54, 50-54.
- Broverman, D. M. Normative and ipsative measurement in psychology. *Psychological Review*, 1962, 69, 295-305.
- Cattell, R. B. Psychological measurement: Normative, ipsative, - interactive. *Psychological Review*, 1944, 51, 292-303.
- Dixon, W. J. *Biomedical computer programs*. Berkeley: University of California Press, 1973.
- Gleser, G. C., & Sacks, M. Ego defenses and reaction to stress: A validation study of the Defense Mechanisms Inventory. *Journal of Consulting and Clinical Psychology*, 1973, 40, 181-187.
- Gleser, G. C., & Ihilevich, D. An objective instrument for measuring defense mechanisms. *Journal of Consulting and Clinical Psychology*, 1969, 33, 51-60.
- Gur, R. E., & Gur, R. C. Defense mechanisms, psychosomatic symptomatology, and conjugative lateral eye movements. *Journal of Consulting and Clinical Psychology*, 1975, 43, 416-420.
- Klein, R. F., Gonen, J. Y., & Smith, C. M. Psychogenic purpura in a man. *Psychosomatic Medicine*, 1975, 37, 41-49.
- Scholz, J. A. Defense styles in suicide attempters. *Journal of Consulting and Clinical Psychology*, 1973, 41, 70-73.
- Warren, J. R. Birth order and social behavior. *Psychological Bulletin*, 1966, 65, 38-39.
- Weissman, H. N., Ritter, K., & Gordon, R. M. Reliability study of the Defense Mechanisms Inventory. *Psychological Reports*, 1971, 29, 1237-1238.

Received December 14, 1977 ■

Blood Nicotine and Carboxyhemoglobin Levels After Rapid-Smoking Aversion Therapy

M. A. H. Russell, Martin Raw, C. Taylor,
C. Feyerabend, and Y. Saloojee
University of London, London, England

Blood nicotine and carboxyhemoglobin (COHb) levels after rapid smoking were studied in 15 smokers. Blood nicotine averaged 48.1 ng/ml after rapid smoking compared to 32.4 ng/ml after normal smoking ($p < .002$), and COHb levels averaged 12.1% and 8.9%, respectively ($p < .001$). Normal smoking levels of 92 smokers in other studies averaged around 30 ng/ml nicotine and 8.2 to 8.5% COHb. There was no evidence that the degree of nicotine and carbon monoxide intoxication produced during rapid smoking had any relation to the reduction in the desire to smoke immediately after the session or to the decrease in cigarette consumption on the following day. The potential risks of rapid smoking are discussed. It is suggested that these risks might be reduced by using a beta adrenergic blocker and that the procedure could be made completely safe, possibly without loss of treatment effect, if subjects were instructed not to inhale.

Among the many aversive techniques devised for the treatment of dependent cigarette smokers, none have so far shown a consistent specific effect when subjected to controlled trial or replication by other research workers. A possible exception is a method of rapid smoking developed by Lichtenstein and his colleagues (Lichtenstein & Danahar, 1976; Lichtenstein, Harris, Birchler, Wahl, & Schmahl, 1973). The procedure involves subjects smoking their usual brand of cigarettes in a rapid and continuous manner, inhaling one puff every 6 sec until no further smoking can be tolerated. It is not clear how much the aversiveness and the subjects' tolerance limits are determined by nicotine intoxication or by local irritation of the mouth and respiratory tract.

Hauser (1974) has pointed out that excessive intake of nicotine could induce cardiac arrhythmia, especially in subjects with

coronary heart disease. Screening to completely exclude coronary heart disease is virtually impossible, and a normal resting (or even posteffort) electrocardiograph is no guarantee of freedom from such disease. Despite Hauser's plea for caution, "thousands of smokers have undergone rapid smoking, many of them in a commercial program operated by Schick Laboratories" (Danahar, Lichtenstein, & Sullivan, 1976, p. 556).

In a therapeutic situation, a degree of risk is justified, provided that it is necessary to offset an even greater risk. Assessment of these risks requires consideration of both severity and probability of the potential consequences of having treatment and going without it. Clearly, therapy should go ahead only if the balance is in the subjects' favor. Furthermore, the assessment should be made in consultation with the subjects, for it is their own evaluation of the severity of consequences that should count. The therapist's role in a decision that is not of clear and obvious overall benefit to the subject should be confined to providing information on the nature and probability of possible hazards.

At present, information on rapid smoking

The authors are grateful to the Medical Research Council and the Department of Health and Social Security for financial support.

Requests for reprints should be sent to M. A. H. Russell, Addiction Research Unit, Institute of Psychiatry, Maudsley Hospital, London, S.E. 5, United Kingdom.

is deficient not only with respect to the degree of hazard, but some doubt also exists as to whether there is any specific benefit. Apart from the numerous other factors in the treatment situation (attention placebo, self-monitoring, persuasion from the therapist etc.), the specific contribution of rapid smoking has been small and in some cases only transient (Lando, 1975). Indeed, Lichtenstein himself in his latest appraisal modestly concluded that "it must be acknowledged that the magnitude of the rapid smoking effect does not appear to be large . . . the interpersonal/persuasive aspects of the treatment setting are a significant source of variance" (Lichtenstein & Danaher, 1976, p. 99).

Of the many toxins present in tobacco smoke, excessive absorption of nicotine and carbon monoxide (CO) represent the greatest potential hazards of rapid smoking, for it is the acute toxic effects that are relevant to this situation rather than the long-term hazards such as lung cancer. Aronow (1976) has reviewed the literature on the effects of nicotine and CO on coronary heart disease. At levels produced by normal smoking, the effects of CO and nicotine interact in such a way as to increase the risk of a sudden heart attack and also to increase the risk of sudden death in the event of such an attack. The interaction is complex but consists basically of the setting up of interrelated vicious cycles. For example, nicotine increases the oxygen requirement of heart muscle, whereas carboxyhemoglobin (COHb) reduces its availability. COHb also impairs the pumping power and efficiency of the heart muscle, leading to further deterioration in its oxygen supply. When the heart muscle is suffering from lack of oxygen, both nicotine and COHb reduce the ventricular fibrillation threshold or, in other words, increase the risk of developing a fatal loss of rhythm.

Few studies have attempted to estimate the hazards of rapid smoking. One study showed that blood COHb, an index of CO absorption, increased from an average of 4.2% before to 7.3% after rapid smoking (Dawley, Ellithorpe, & Tretola, 1976). Un-

fortunately, no details are given as to the method of rapid smoking used, and crucial data such as rate, duration, and number of cigarettes smoked during the session are not mentioned. Another study by Lichtenstein and his colleagues (Danaher et al., 1976) showed that rapid smoking produced a significantly greater rise in heart rate than normal smoking, but they also found that CO absorption was not significantly greater. However, their method of COHb estimation was somewhat archaic, and they conceded the need for further study using more sophisticated measures. More recently, Horan, Hackett, Nicholas, Linberg, Stone, & Lukaski (1977) have reported the occurrence of electrocardiographic abnormalities during rapid smoking, and they suggested that these may be due to nicotine intoxication (Horan, Linberg, & Hackett, 1977). Indeed, they commented that "normative data on the amount of nicotine absorbed by subjects during rapid smoking are an urgent research priority" (Horan, Linberg, and Hackett, 1977, p. 346).

We are conducting an evaluative trial of rapid smoking among subjects attending a tobacco withdrawal clinic. Full analysis of our findings is awaiting long-term follow-up. However, in those subjects randomly assigned to rapid smoking, we have measured blood nicotine and COHb levels after normal smoking and after their first session of rapid smoking. We have also been able to compare the levels produced in these subjects by rapid smoking with the average normal-smoking levels of subjects from a number of previous studies. Finally, to see whether nicotine intoxication contributes to the therapeutic effect, we have examined the relation between the degree of excessive nicotine intake during rapid smoking and the reduction in cigarette consumption on the day after the first session.

These findings are presented here, and the implications for the possible hazards of rapid smoking are discussed. No other study has measured nicotine intake from rapid smoking, and excessive nicotine intake is the main potential hazard. It was indeed the main concern of Hauser (1974).

Method

Subjects

The subjects were 5 men and 10 women who attended the Smokers Withdrawal Clinic at the Maudsley Hospital, London, England, and who volunteered to take part in a clinical trial of some "new behavioral treatments." The trial in question was a straight comparative study of rapid smoking, cue exposure, and simple support. (This study is still in progress at the stage of long-term follow-up.) Subjects were excluded if they had a clinical or family history suggesting the possibility of coronary heart disease. All men over 40 years old and women over 50 were excluded even in the absence of such evidence. Of the 17 eligible subjects who were randomly assigned to rapid smoking, complete blood data were missing for 2; 1 had a needle phobia, and in the case of the other, a specimen tube was broken in the laboratory.

Procedure

Venous blood samples were taken on the day of the first session of a course of rapid-smoking treatment. The sessions took place in the late afternoon (between 4 p.m. and 6 p.m.), and subjects were instructed to smoke normally until the start of treatment. On arrival at the clinic, they were asked to smoke a cigarette in their usual way, and a blood sample (normal-smoking sample) was then taken 2 minutes after the cigarette was completed. The blood nicotine concentration falls rapidly from a peak just after a cigarette. In all of our studies, we have used a time interval of 2 minutes after a cigarette to approximate to the peak level of nicotine. The rapid-smoking session was started about 15–20 minutes after the normal-smoking sample, and a second blood sample was taken as close as possible to 2 minutes after completion of the last trial of the session. As the sessions were given in small groups of four or five subjects, it was not always possible to take the blood within 2 minutes of discarding the last cigarette of the session. For one or two subjects, it was taken at 4–6 minutes. The error from such delay would tend to make the rapid-smoking nicotine values appear slightly lower than their true levels.

To have reversed the order of sampling in half of the subjects by taking the normal-smoking sample after the rapid-smoking sample might have been methodologically more correct, but a cigarette taken even 30 minutes after a rapid-smoking session would not have been quite "normal." Besides, the subjects had been smoking normally all day, and the cigarette smoked for the "normal-smoking sample" would have had little extra effect on the nicotine level of the "rapid-smoking sample" (Russell, Feyerabend, & Cole, 1976).

The rapid smoking was conducted as follows: The subjects were instructed to puff and inhale deeply every 6 sec and to continue for as long as

possible, even if it meant lighting a second or third cigarette. This was followed by a 5-min rest, after which the sequence was repeated for up to three trials if necessary. Many subjects felt unable to face a third trial.

The desire to smoke was rated on a 7-point scale before the first and after the last trial. The ratings were as follows: 0 = none, 1 = very slight, 2 = slight, 3 = moderate, 4 = fairly strong, 5 = strong, 6 = very strong.

Blood Analysis

Blood samples were analysed for COHb using an IL 182 CO-oximeter (Russell, Cole, & Brown, 1973) and for nicotine using gas chromatography (Feyerabend, Levitt, & Russell, 1975). The laboratory workers who did the analysis were unaware of the design or purpose of the study.

Results

Individual data are shown in addition to the means (Table 1). When individual risk is involved, it is relevant if only 1 case in 20 shows an excessive or unusual response.

Rapid Smoking in Practice

The amount of rapid smoking undertaken was subjectively determined by the individual's own tolerance. Seven subjects had three trials and seven had two trials, but Subject 15 (Table 1) could manage only one trial. On average, 3.3 cigarettes were smoked during the session at a rate of 2.5 min per cigarette. Thus, at 1 puff every 6 sec, a mean of 25 puffs was obtained from each cigarette, which is at least double the number taken during normal smoking and by the standardized puffing of the smoking machine at routine analysis of tar and nicotine yields (Rothwell & Grant, 1974). The total duration of smoking during the session averaged 8.3 min, but it varied widely between subjects, ranging from 2 min (Subject 8) to 14.25 min (Subject 4). There was a tendency for the amount of rapid smoking tolerated to decrease with each successive trial. The average duration of the trials were 4.6, 3.0, and 2.1 min for Trials 1–3, respectively, and the average number of cigarettes smoked per trial were 1.7, 1.4, and .8, respectively. The tolerance limits were determined by irritation to the mouth and throat (six subjects) or by

Table 1

Blood Nicotine and Carboxyhemoglobin (COHb) Levels After Rapid Smoking

Subjects	Initial daily cigarette consumption	Nicotine yield of cigarettes (mg)	No. cigarettes smoked in session	COHb (%)		Plasma nicotine (ng/ml)	
				Before rapid smoking	After rapid smoking	After normal smoking	After rapid smoking
Males							
1	26 (45)	—	3.5	2.8	5.3	49.6	64.0
2	34 (27)	1.0	4.3	12.1	15.8	16.2	58.3
3 ^a	72 (60)	1.9	3.5	11.9	14.2	41.6	70.0
4 ^a	21 (25)	.7	4.5	9.0	13.1	25.7	54.5
5	17 (20)	1.4	2.5	9.2	12.6	25.3	29.3
<i>M</i>	34.0 (35.4)	1.3	3.7	9.0	12.2	31.7	55.2
$\pm SD$	± 22.2 (16.7)	$\pm .5$	$\pm .8$	± 3.8	± 4.0	± 13.6	± 15.6
Females							
6	28 (30)	.7	4.0	8.0	11.5	29.4	62.8
7	21 (30)	1.3	2.8	7.7	10.2	32.6	48.3
8	14 (18)	1.2	1.0	3.9	4.8	12.9	10.8
9 ^a	34 (40)	.7	3.3	7.7	11.0	29.8	53.6
10 ^a	33 (35)	.7	2.6	8.0	10.4	29.8	29.1
11 ^a	28 (40)	1.3	4.5	10.6	14.5	37.2	44.4
12	36 (45)	1.2	3.5	10.7	14.4	42.5	58.9
13 ^a	30 (45)	1.2	4.5	10.2	14.4	49.2	38.2
14 ^a	18 (20)	.9	4.0	9.4	13.8	19.2	48.1
15	24 (23)	1.3	1.5	12.9	15.5	45.5	51.0
<i>M</i>	26.6 (32.6)	1.1	3.2	8.9	12.1	32.8	44.5
$\pm SD$	± 7.2 (10.0)	$\pm .3$	± 1.2	± 2.4	± 3.2	± 11.3	± 15.3
Total	29.1 (33.5)	1.1	3.3	8.9	12.1	32.4	48.1
<i>M</i> \pm <i>SD</i>	± 13.7 (12.1)	$\pm .3$	± 1.1	± 2.8	± 3.4	± 11.6	± 15.7

Note. The data for daily cigarette consumption are the self-recorded levels, with the initial levels claimed in parentheses. Subject 1 rolled his own cigarettes, so that the nicotine yield was unknown. None of the differences between men and women were statistically significant.

* These subjects tolerated three trials during the session; the others all had two trials except Subject 15, who managed only one trial.

nausea (nine subjects), but no subject vomited. The experience of nausea was not significantly associated with either relative or absolute increase in plasma nicotine level after rapid smoking.

Carbon Monoxide Intake

COHb levels before and after rapid smoking are shown in Table 1. In view of the slow removal of CO from the body, especially under sedentary conditions, the decrease over the 15- to 20-minute period between taking the "before" sample and starting rapid smoking would have been small (< .5% COHb; Russell, Wilson, Cole, Idle, & Feyerabend, 1973). However, this means that the full increase attributable to rapid smoking

would have been a little more than the average increase per subject of 3.2 ($SD = .9$) % COHb. As expected, the difference between the mean COHb before and after rapid smoking was highly significant, $t(14) = 12.99$, $p < .0001$. The increase in COHb was related to the number of cigarettes smoked in the session, $r(14) = .81$, $p < .01$, and to the total duration of smoking during the session, $r(14) = .58$, $p < .05$, but it was not significantly greater in those who tolerated three compared to two trials during the session (3.5 vs. 2.9% COHb), $t(11) = 1.2$, *ns*.

Plasma Nicotine Levels

The comparison of plasma nicotine concentrations after rapid smoking with peak

levels obtained from normal smoking are shown in Table 1. The average of 48.1 ng/ml after rapid smoking is significantly higher than the normal smoking mean of 32.4 ng/ml, $t(14) = 4.01$, $p < .002$. Plasma nicotine decreases rapidly over the first 10–15 minutes after completing a cigarette (Isaac & Rand, 1972), so that the levels after rapid smoking would have been even higher had all the blood samples been taken within 2 minutes of completing the last trial. It is probable that a slight delay in taking the blood accounted for the few subjects whose rapid-smoking levels did not substantially exceed the normal-smoking level (e.g., Subjects 8, 10, and 13). However, no record was kept as to which subjects did experience such delay. The average relative excess of plasma nicotine after rapid smoking was 48% above the normal-smoking level (Table 1).

The plasma nicotine after rapid smoking was related to the number of cigarettes smoked in the session, $r(14) = .55$, $p < .05$, and to a lesser extent with total duration of smoking during the session ($r = .44$, *ns*). There was no significant difference in the levels obtained by those who tolerated three versus two trials, for whom the means were 48.3 and 47.5 ng/ml, respectively. Plasma nicotine levels after normal smoking and rapid smoking were not significantly correlated ($r = .42$, *ns*). The association between the subjects' plasma nicotine levels and their usual cigarette consumption was statistically significant for rapid smoking, $r(14) = .56$, $p < .05$, but not for normal smoking ($r = .37$, *ns*). The nicotine yield of the cigarette smoked seemed to have little effect on the plasma nicotine level produced ($r = .40$, *ns*, for normal smoking; and $r = .06$ for rapid smoking).

Heart Rate

Pulse rate was counted for only four subjects. However, it seemed clear that the increase was produced by the first rapid-smoking trial and that the second trial produced no further increase. The four subjects had a mean rate of 88 per minute before start-

Table 2

Lack of Relation of Treatment Effects of the First Session to Rapid-Smoking Variables and Plasma Nicotine and Carboxyhemoglobin (COHb) Increase During Rapid Smoking

Variable	Reduction in desire to smoke	% reduction in cigarette consumption on day after the session
<i>M</i> for subjects who had		
2 trials ^a	3.7	75.4
3 trials ^a	2.7	55.8
<i>t</i>	1.04	.99
Product-moment correlations ^b		
No. cigarettes smoked in session	.33	.27
Time taken to smoke each cigarette	-.21	-.38
Total smoking time in session	.13	-.04
Increase in COHb%	.29	.22
Blood nicotine after rapid smoking	.30	.28
Ratio of rapid- to normal-smoking blood nicotine	.21	.29

Note. One subject (15) tolerated only one trial. For a sample size of 15, a 5% level of significance requires a correlation of .51 or more. The reduction in the rating of desire to smoke over the course of the session correlated .60 with the percentage of reduction in cigarette consumption on the following day.

^a $n = 7$.

^b $n = 15$.

ing the session. This increased to 106.5 after the first and 104.7 after the second trial.

Treatment Response to Rapid Smoking

A full assessment of the value of rapid smoking as a treatment for smokers will be reported separately after a 1-year follow-up, but two potentially relevant variables for treatment outcome deserve mention here because of their lack of significant association with the amount of rapid smoking in the session or the degree of nicotine and CO intake (Table 2).

Subjective ratings of desire to smoke. The average rating of 3.7 ($SD = 1.7$) before the session decreased to .6 ($SD = 1.0$) just after the last trial, $t(14) = 6.87$, $p < .001$.

Table 3

Means \pm Standard Deviations for Blood Nicotine and Carboxyhemoglobin Levels After Rapid Smoking Compared to Normal Smoking Data From Other Studies

Study	No. subjects	Usual daily cigarette consumption	COHb (%)	Plasma nicotine (ng/ml)
Other				
Russell, Wilson, Patel, Cole, and Feyerabend (1973); Russell, Wilson, Patel, Feyerabend, and Cole (1975)	10	27.2 \pm 6.9	8.2 \pm 2.2	30.1 \pm 10.7
Russell, Wilson, Feyerabend, and Cole (1976)	43	33.3 \pm 17.1	8.5 \pm 2.6	30.1 \pm 12.5
Russell, Sutton, Feyerabend, Cole, and Saloojee (1977)	21	33.1 \pm 11.5	8.2 \pm 2.7	31.5 \pm 12.4
Sutton, Feyerabend, Cole, and Russell (1978)	18	28.8 \pm 9.8	8.5 \pm 2.1	29.0 \pm 17.3
Present				
Normal smoking	15	29.1 \pm 13.7	8.9 \pm 2.8	32.4 \pm 11.6
Rapid smoking			12.1 \pm 3.4	48.1 (\pm 15.7)

Note. COHb = carboxyhemoglobin. Blood samples were all taken in the afternoon approximately 2 minutes after completing a cigarette. With the exception of rapid smoking, all subjects had spent the day smoking their usual brand of cigarettes in their usual manner.

The subjects had been smoking normally and had not been deprived before the session, so the initial rating was not very high.

Self-recorded cigarette consumption. On the day after the rapid-smoking session, cigarette consumption decreased on average to 32.7% ($SD = 34.4$) of the baseline level before treatment. Only 4 of the 15 subjects failed to achieve a reduction of at least 50%, and 3 did not smoke at all.

Discussion

There seems little doubt from our data that rapid smoking can give rise to blood levels of nicotine and COHb that greatly exceed those produced by the normal smoking of very heavy smokers. Our data from five different studies, comprised of 107 smokers, are consistent (Table 3). They show that our subjects have been heavy smokers with an average consumption of about 30 cigarettes per day compared to a national average in Britain of 22 per day for men and 16 per day for women (Lee, 1976). With normal smoking, COHb levels taken during the afternoon approximately 2 minutes after a cigarette averaged 8.0%–8.5%, whereas plasma nicotine averaged about 30.0 ng/ml. In contrast, rapid smoking produced an average

COHb of 12.1% and a plasma nicotine concentration of 48.1 ng/ml. The average increase in % COHb during a rapid-smoking session was 3.2, which is the same as the increase found by Dawley et al. (1976).

Since nicotine and CO intake from normal smoking constitute a risk for people with coronary heart disease (Aronow, 1976), the risk must be greater during rapid smoking. Unfortunately, the magnitude of the risk is not easily assessed. In the first place, no one has yet attempted to quantify the short-term risk of say 1 day of normal smoking to subjects with severe coronary heart disease, let alone for subjects without evidence of such disease. Had our data shown that plasma nicotine and COHb levels after rapid smoking were no higher than after normal smoking, we could have concluded that the risks did not increase. This was clearly not the case, so that one is beholden to make some attempt to assess the risks.

On the average, plasma nicotine levels after rapid smoking were about 50% higher and COHb levels 36% higher than after normal smoking. It is unlikely that the hazards increase in simple linear relation to the blood levels. They probably accelerate positively from some point. This probably also occurs with the adverse synergistic interac-

tion between nicotine and COHb on cardiac function (Aronow, 1976). It would, therefore, be reasonable to assume that the risks are probably increased far more than is indicated by simple comparison of blood levels. Furthermore, assessment of risk should not be confined to consideration of averages. Some individuals had exceptionally high blood nicotine and COHb levels after rapid smoking (Table 1). Subjects 2 and 14 had nicotine levels that were, respectively, 260% and 150% higher than their normal smoking level.

In a recent article, Horan, Linberg, and Hackett (1977) attempted to assess the extent to which nicotine poisoning occurs in rapid smoking. Most of their information was derived from an outdated edition of a well-known textbook of pharmacology (Goodman & Gilman, 3rd ed., 1965, rather than Goodman & Gilman, 5th ed. in 1975), and their article contains several errors and misinterpretations. They stated, for example, that a single cigarette yields about 6-8 mg of nicotine, yet the average nicotine yield of cigarettes smoked in the United States in 1975 was 1.2 mg, and the strongest brand on the market yielded only 2.1 mg (Owen, 1976). Even the hardest puffing smoker would find it difficult to abstract more than about three times the standard machine-smoked yield from a cigarette. Horan, Linberg, and Hackett also assumed that increasing the puff rate would proportionally increase the number of puffs, and hence the nicotine dose, obtained from a single cigarette (puff volume remaining the same). This is simply not true. It is doubtful whether a 10-fold increase in puff rate (from 1 per min in the case of standard machine smoking to 1 per 6 sec as in rapid smoking) would increase the number of puffs obtained from a cigarette by more than a factor of three. The average of 25 puffs per cigarette obtained by our subjects with rapid smoking represents a factor of approximately 2.5, and this was probably due more to a reduction in puff volume than the increase in puff rate. Horan, Linberg, and Hackett seemed not to appreciate that increases in puff rate produce higher levels of plasma nicotine (but not COHb), not so much by increasing the nicot-

tine dose per cigarette as by increasing the rate at which the dose is taken.

In our view, the incidence of cardiac complications is of greater significance than the occurrence of nicotine poisoning or intoxication. Nicotine intoxication per se is fairly rapidly reversible. Indeed, the pharmacological effects of nicotine that are evident in many smokers after normal smoking are manifestations of a degree of nicotine intoxication. So-called "green-tobacco sickness," an occupational illness of tobacco harvesters, is a form of nicotine intoxication due to absorption through the skin during tobacco cropping (Gehlbach et al., 1975). The symptoms include nausea, vomiting, dizziness, and prostration, far exceeding those usually encountered after rapid smoking. Mortality and long-term effects of green tobacco sickness have not been documented, but during the 1973 harvesting season an estimated 9% of North Carolina's 60,000 tobacco growers reported such illness among their workers (Gehlbach et al., 1975). As far as rapid smoking is concerned, the severity of manifest nicotine intoxication is relevant only insofar as it is an index of nicotine dosage, and hence the risk of serious cardiac complications that could persist after nicotine intoxication has passed. However, the relation of cardiac complications to signs of nicotine intoxication is unlikely to be very close. One reason is that the incidence of cardiac complications depends on the extent of underlying coronary heart disease. Another reason is that the tolerance mechanisms probably differ for the effects of nicotine on the brain as opposed to the heart and circulation. Finally, in our subjects the experience of nausea after rapid smoking was not significantly associated with high nicotine intake.

In view of the complexities, it is virtually impossible to make any meaningful quantitative estimate of the risks of rapid smoking on the basis of the blood level data alone. Since the main concern is for the cardiac complications, the most relevant approach to assessing the risks should certainly include careful electrocardiograph analysis during and after rapid smoking (Horan, Hackett, Nicholas, Linberg, Stone, & Lukaski, 1977). Ironically, though it is usual to make some esti-

mate of risk before widespread use of a new treatment approach, the most reassuring evidence of the relative safety of rapid smoking is its use for some years with "thousands of smokers" (Danaher et al., 1976) without apparent mishap.

The most careful screening cannot exclude the presence of even quite advanced coronary artery disease. Rapid smoking must, therefore, involve a greater risk to older subjects, 35 and up, especially if they are men. Besides careful screening, any risk of cardiac arrhythmia due to rapid smoking could be further reduced by the administration of a beta adrenergic blocker such as oxprenolol. This drug has been shown to reduce the increase in heart rate due to normal smoking (Carruthers, 1976). This approach might be especially relevant to those cases who are at greater risk but for whom it is particularly important to give up smoking.

As mentioned above, a degree of risk is more justified if rapid smoking can be shown to be the most likely way to enable a person to stop smoking. It was not the purpose of this article to evaluate rapid smoking as a treatment for smokers. Nevertheless, the fact that neither COHb increase during rapid smoking nor plasma nicotine levels after rapid smoking correlated significantly with reduction in cigarette consumption, on the one hand, or lowering of the desire to smoke, on the other (Table 2), suggests that excessive intake of nicotine and CO may not be necessary elements of any treatment effect. Furthermore, it seems that there is little to be gained by pressing subjects to undertake more than two rapid-smoking trials per session, since subjects who tolerate only two trials show, if anything, a greater reduction in both desire to smoke and cigarette consumption than do those who go on to a third trial. This is not explained by the possibility that heavier smokers might be more likely to tolerate three trials, since mean cigarette consumption did not differ significantly in the two groups ($t = 1.15$, ns). Of course, the possibility remains that two-trial subjects might have done better if pushed to do a third, or that three-trial subjects would have shown even less change after only two trials.

It could be argued that the lack of correla-

tion between early treatment outcome and nicotine and CO intake during rapid smoking was due to differences in individual tolerance. For example, a nicotine level of 30 ng/ml might be highly aversive to one subject but relative deprivation to another whose regular smoking level was around 50 ng/ml. However, this is an unlikely explanation, since the ratio of rapid-smoking to normal-smoking plasma nicotine level did not correlate significantly with reduction in desire to smoke after the session or with the reduction in cigarette consumption on the following day (Table 2).

Our data on heart rate increase (21% after the first trial and 19% after the second) show greater changes than the 9% increase reported by Hynd, O'Neal, and Severson (1976) but are in accord with the more substantial, controlled study by Danaher et al. (1976). A larger effect on heart rate could be expected when rapid smoking follows a period of abstinence.

It is concluded that rapid smoking can produce excessively high blood nicotine and COHb levels and that this constitutes a risk to all but the younger smoker. On the data available at present, the level of risk is impossible to calculate in quantitative terms. It could probably be reduced by a beta adrenergic blocking drug. Since excessive nicotine and CO intake may not be necessary for a treatment effect, it is suggested that the procedure could be made completely safe and possibly no less effective if subjects are instructed to take the smoke into the throat but to avoid inhaling it into the lungs, for it is only by inhalation that dangerously high levels are produced.

References

- Aronow, W. S. A critical review of the effect of nicotine and carbon monoxide on coronary heart disease. *World Smoking and Health*, 1976, 1, 20-24.
- Carruthers, M. Modification of the noradrenaline related effects of smoking by beta-blockade. *Psychological Medicine*, 1976, 6, 251-256.
- Danaher, B. G., Lichtenstein, E., & Sullivan, J. M. Comparative effects of rapid and normal smoking on heart rate and carboxyhemoglobin. *Journal of Consulting and Clinical Psychology*, 1976, 44, 556-563.

- Dawley, H. H., Ellithorpe, D. B., & Tretola, R. Aversive smoking: Carboxyhaemoglobin levels before and after rapid smoking. *Journal of Behavior Therapy and Experimental Psychiatry*, 1976, 7, 13-15.
- Feyerabend, C., Levitt, T., & Russell, M. A. H. A rapid gas-liquid chromatographic estimation of nicotine in biological fluids. *Journal of Pharmacy and Pharmacology*, 1975, 27, 434-436.
- Gehlbach, S. H., et al. Nicotine absorption by workers harvesting green tobacco. *Lancet*, 1975, 1, 478-480.
- Goodman, L. S., & Gilman, A. (Eds.). *The pharmacological basis of therapeutics* (3rd ed.). New York: MacMillan, 1965.
- Goodman, L. S., & Gilman, A. (Eds.). *The pharmacological basis of therapeutics* (5th ed.). New York: MacMillan, 1975.
- Hauser, R. Rapid smoking as a technique of behavior modification: Caution in selection of subjects. *Journal of Consulting and Clinical Psychology*, 1974, 42, 625-626.
- Horan, J. J., Hackett, G., Nicholas, W. C., Linberg, S. E., Stone, C. I., & Lukaski, H. C. Rapid smoking: A cautionary note. *Journal of Consulting and Clinical Psychology*, 1977, 45, 341-343.
- Horan, J. J., Linberg, S. E., & Hackett, G. Nicotine poisoning and rapid smoking. *Journal of Consulting and Clinical Psychology*, 1977, 45, 344-375.
- Hynd, G. W., O'Neal, M., & Severson, H. H. Cardiovascular stress during the rapid-smoking procedure. *Psychological Reports*, 1976, 39, 371-375.
- Isaac, P. F., & Rand, M. J. Cigarette smoking and plasma levels of nicotine. *Nature*, 1972, 236, 308-310.
- Lando, H. A. A comparison of excessive and rapid smoking in the modification of chronic smoking behavior. *Journal of Consulting and Clinical Psychology*, 1975, 43, 350-355.
- Lee, P. N. (Ed.). *Statistics of smoking in the United Kingdom* (7th ed.). London: Tobacco Research Council, 1976.
- Lichtenstein, E., & Danaher, B. G. Modification of smoking behavior: A critical analysis of theory, research and practice. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 3). New York: Academic Press, 1976.
- Lichtenstein, E., Harris, D., Birchler, G., Wahl, J., & Schmahl, D. Comparison of rapid smoking, warm smoky air, and attention placebo in the modification of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1973, 40, 92-98.
- Owen, T. B. Tar and nicotine from U.S. cigarettes: Trends over the past twenty years. In E. L. Wynder, D. Hoffman, & G. B. Gori (Eds.), *Proceedings of the 3rd World Conference on Smoking and Health* (Vol. 1): *Modifying the risk for the smoker*. Washington, D.C.: U.S. Department of Health, Education and Welfare, 1976.
- Rothwell, K., & Grant, C. A. *Standard methods for the analysis of tobacco smoke* (Research Paper 11, 2nd ed.). London: Tobacco Research Council, 1974.
- Russell, M. A. H., Cole, P. V., & Brown, E. Absorption by non-smokers of carbon monoxide from room-air polluted by tobacco smoke, *Lancet*, 1973, 1, 576-579.
- Russell, M. A. H., Feyerabend, C., & Cole, P. V. Plasma nicotine levels after cigarette smoking and chewing nicotine gum. *British Medical Journal*, 1976, 1, 1043-1046.
- Russell, M. A. H., Sutton, S. R., Feyerabend, C., Cole, P. V., & Saloojee, Y. Nicotine chewing gum as a substitute for smoking. *British Medical Journal*, 1977, 1, 1060-1063.
- Russell, M. A. H., Wilson, C., Cole, P. V., Idle, M., & Feyerabend, C. Comparison of increases in carboxyhaemoglobin after smoking "extra-mild" and "non-mild" cigarettes. *Lancet*, 1973, 2, 687-690.
- Russell, M. A. H., Wilson, C., Feyerabend, C., & Cole, P. V. Effect of nicotine chewing gum on smoking behaviour and as an aid to cigarette withdrawal. *British Medical Journal*, 1976, 2, 391-395.
- Russell, M. A. H., Wilson, C., Patel, U. A., Cole, P. V., & Feyerabend, C. Comparison of effect on tobacco consumption and carbon monoxide absorption of changing to high and low nicotine cigarettes. *British Medical Journal*, 1973, 4, 512-516.
- Russell, M. A. H., Wilson, C., Patel, U. A., Feyerabend, C., & Cole, P. V. Plasma nicotine levels after smoking cigarettes with high, medium and low nicotine yields. *British Medical Journal*, 1975, 2, 414-416.
- Sutton, S. R., Feyerabend, C., Cole, P. V., & Russell, M. A. H. Adjustment of smokers to dilution of tobacco smoke by ventilated cigarette holders. *Clinical Pharmacology and Therapeutics*, 1978, 24, 395-405.

Received December 15, 1977 ■

The Hand Dynamometer as a Neuropsychological Measure

Carl B. Dodrill

Department of Neurological Surgery
University of Washington School of Medicine

The sensitivity of the hand dynamometer to the presence of brain damage and to its lateralization was evaluated and compared with that of the Tapping Test and the Tactual Performance Test. Four groups of 25 subjects each were studied (control, right-hemisphere damage, left-hemisphere damage, and bilateral damage). Measures of performance on each test included those of each hand taken separately as well as their sum. To identify the lateralization of brain lesions, a method was developed that used the control group as a basis for comparison and that simultaneously considered the relative performances of each hand on each task. All test variables discriminated between the control and brain-damaged groups at high levels of statistical significance. Furthermore, the dynamometer discriminated between these groups as well as did the Tapping Test and Tactual Performance Test. Finally, the dynamometer correctly identified the lateralization of brain lesions in more instances than either of the other tests. It is concluded that the hand dynamometer is a neuropsychological measure of considerable promise.

Many years ago, Halstead (1947) demonstrated that assessment of voluntary motor movement could be useful in evaluating the integrity of brain functions using such measures as the Tapping Test and the Tactual Performance Test. Reitan (1966) expanded the use of these measures by demonstrating that differences in performance between the two hands are related to the relative functioning capabilities of the two cerebral hemispheres. Thus, by examining level of performance and by comparing the two sides of the body, these tests of motor speed and

agility were established as reliable indicators of the integrity of brain functions.

Clinicians have realized that intensity or strength of voluntary motor activity might also be a reliable indicator of brain functions, and many use some strength-of-grip measure in neuropsychological assessment. Such use has led to an establishment of its clinical value as well as to a listing of the dynamometer by Reitan and Davison (1974) as a neuropsychological measure. On a research basis, Reitan (1974) demonstrated that the strength of grip of young brain-damaged and nonneurological children (ages 5-8) differs only slightly. Boll (1974), in working with older children (ages 9-14), found much more striking differences. No parallel studies have been done with adults that have directly compared brain-damaged persons with nonneurological controls, and none have evaluated the dynamometer with respect to the correct placement of lateralized lesions. The present study addresses these areas and evaluates the utility of the hand dynamometer in comparison with two other better established neuropsychological measures (Tapping Test, Tactual Performance Test).

This project was supported by National Institutes of Health Contracts N01-NS-0-2281 and N01-NS-6-2341 and by National Institutes of Health Grant NS 04053 awarded by the National Institute of Neurological and Communicative Disorders and Stroke, U.S. Public Health Service, Department of Health, Education, and Welfare.

The author gratefully acknowledges the contribution of Ralph M. Reitan to this study by providing the data for the neurological groups.

Requests for reprints should be sent to Carl B. Dodrill, who is also at the Epilepsy Center (ZA-50), Harborview Medical Center, 325 Ninth Street, Seattle, Washington 98104.

Method

Subjects

Four groups of adults (ages 15 and over) were formed, with each group consisting of 25 persons. Subjects in the control group had negative neurological histories. They had never had any disease that might have affected the nervous system (meningitis, encephalitis, polio, diabetes, rheumatic fever, scarlet fever, etc.), and they had no histories of high fever, partial drowning, exposure to gas, heat exhaustion, fainting spells, or head trauma. They were recruited from a variety of community resources including churches, schools, and employment agencies.

Three groups of brain-damaged persons were selected on the basis of the primary location of brain damage (right hemisphere, left hemisphere, both cerebral hemispheres). In each group, there were 5 individuals with intrinsic brain tumors, 11 with a history of head trauma, and 9 with cerebral vascular problems. Neurological diagnoses were established by anamnestic information, angiography, pneumoencephalography, electroencephalography, skull x-rays, neurosurgical findings, and autopsy.

Across all groups, a subject-by-subject matching procedure was maintained for the variables of sex (there were 20 males and 5 females in each group), race (all subjects were Caucasian), and handedness (all subjects were right-handed). Within the brain-damaged groups, the subject-by-subject matching procedure included the general type of neurological difficulty (neoplastic, traumatic, vascular). Finally, within each set of 4 persons (1 from each group), matching was completed as closely as possible for age and years of formal education, with the result that each group averaged approximately 41.14 years of age and 10.68 years of education.

As part of their neuropsychological evaluations, all subjects were administered the dynamometer, the Tapping Test, and the Tactual Performance Test. Attention was given to the exact administrative procedures suggested by Reitan with a strong emphasis on maximal performance. To assess strength of grip, the Smedley Hand Dynamometer was used, which registers strength in kilograms. Two trials were given in alternating fashion for each hand and the beginning with the right (preferred) hand, and the average of the two trials was used as the final score for each hand.

Because the Tactual Performance Test provided a total time score summing all trials (including right hand, left hand, and both hands), summary scores (right plus left) for the Tapping Test and the dynamometer were also provided in addition to the usual scores for each hand alone.

Analyses

To evaluate the discriminability of the tests, univariate analyses of variance were run across all four groups for each test variable, and evaluations of significant differences between groups were assessed by the Newman-Keuls procedure (Winer,

1971). In these analyses, homogeneity of variance was maintained by converting all data to normalized standard scores with a mean of 50 and a standard deviation of 10. Performances on the Tactual Performance Test were considered on a minutes-per-block basis.

The effectiveness of each of the three tests in implicating lateralized damage was assessed using only the subjects with lateralized lesions. The performance by the left (nonpreferred) hand was divided by the performance of the right (preferred) hand so that in each instance a left-to-right comparison in performance could be made with a single score. The mean and standard deviation of this score for the control group were computed, and 1 standard deviation on either side of the mean was arbitrarily selected as the limit of normal performance. The performance of each individual in the right and left brain-damaged groups was then compared with this standard. If the performance for any brain-damaged patient on each of the three measures considered separately indicated that the right hand was not performing as well as would be expected in comparison with the performance of the left hand, the left cerebral hemisphere was considered to be implicated by that measure, and vice versa. Chi-square statistics were applied to the subjects who were classified by this procedure.

Results

The discriminability of each neuropsychological variable considered on a group-by-group basis is given in Table 1. Highly statistically significant differences across the groups were found with respect to every variable, and the control group did better than all brain-damaged groups in every instance. The dynamometer discriminated between the normal and brain-damaged groups as well as did either of the other tasks.

The lateralization data are presented in Table 2. If performance fell within the normal (1 standard deviation) range, neither hemisphere was considered implicated and placement was made in the "neither" group. When one hemisphere or the other was implicated, all three tests classified a majority of individuals correctly, and the dynamometer correctly classified the largest number.

Discussion

The high level of discriminability demonstrated by the dynamometer between normal and brain-damaged subjects was unexpected. It is true that the brain-damaged groups had

Table 1
Data on All Test Variables for All Groups

Test and variable	Control		Right damage		Left damage		Bilateral damage		F
	M	SD	M	SD	M	SD	M	SD	
Dynamometer									
Right hand	48.12 ^{b,c,d}	13.39	33.94 ^a	12.07	31.42 ^a	16.85	36.16 ^a	9.37	7.73
Left hand	44.86 ^{b,c,d}	12.15	21.42 ^{a,c,d}	15.18	37.21 ^{a,b}	12.75	32.44 ^{a,b}	11.83	13.73
Total (right + left)	93.96 ^{b,c,d}	25.12	55.36 ^{a,c,d}	22.24	68.61 ^{a,b}	26.48	68.60 ^{a,b}	19.61	10.43
Tapping									
Right hand	53.44 ^{b,c,d}	6.23	41.76 ^a	9.00	37.00 ^a	17.28	40.36 ^a	11.21	9.64
Left hand	49.60 ^{b,c,d}	5.37	30.72 ^{a,c}	14.30	39.24 ^{a,b,d}	11.50	34.60 ^{a,c}	10.28	17.11
Total (right + left)	103.04 ^{b,c,d}	10.43	72.48 ^a	20.27	76.24 ^a	26.08	75.44 ^a	18.97	11.96
Tactual Performance									
Right hand	.76 ^{b,c,d}	.50	4.42 ^a	5.54	5.70 ^{a,b}	6.03	2.54 ^{a,c}	3.28	12.25
Left hand	.64 ^{b,c,d}	.37	6.83 ^{a,c,d}	6.85	2.77 ^{a,b}	4.69	2.45 ^{a,b}	3.30	8.53
Both hands	.39 ^{b,c,d}	.21	4.24 ^a	5.94	2.27 ^a	4.29	1.59 ^a	3.15	8.13
Total (all trials)	.59 ^{b,c,d}	.32	4.14 ^{a,d}	5.25	2.63 ^a	4.18	1.74 ^{a,b}	2.16	11.07

Note. *F* statistics were computed on the basis of *T* scores. All *F*s were significant at the .001 level. *n* = 25. Superscripts designated groups with statistically different performances (*p* < .01).

^a Control.

^b Right damage.

^c Left damage.

^d Bilateral damage.

unequivocal evidence of cerebral involvement. It is also true that the nonneurological group, consisted of "off the street" individuals rather than the hospital populations usually studied (Halstead, 1947; Reitan, 1955; Vega & Parsons, 1967). These facts may have

Table 2
Numbers of Subjects in the Right- and Left-Damaged Groups Classified According to the Lateralizing Implications of Their Performance

Test and group	Hemisphere implicated			χ^2
	Right	Left	Neither	
Dynamometer				
Right damaged	12	1	12	16.88**
Left damaged	3	13	9	
Tapping				
Right damaged	10	6	9	6.01*
Left damaged	3	7	15	
Tactual Performance				
Right damaged	9	3	13	6.39*
Left damaged	2	7	16	

* *p* < .05.

** *p* < .001.

served to accentuate the general differences between the groups, but they did not give the Tapping Test or the Tactual Performance Test any noticeable edge in discriminant ability over the dynamometer. This was particularly surprising in view of the extreme simplicity of the dynamometer. An incidental observation was that when performances by the preferred hand alone were considered, control subjects outperformed their matched brain-damaged subjects 91% of the time with the dynamometer, 82% of the time with the Tapping Test, and 84% of the time with the Tactual Performance Test. Thus, it appears that the dynamometer does effectively discriminate between normal and brain-damaged adults when consideration is made either on a subject-by-subject or on a group-by-group basis.

The discriminability of the dynamometer may in part relate to the age of the person to whom it is administered. It is of interest to note that Reitan (1974) showed only minimal discrimination between normal and brain-damaged young children with the dynamometer, but Boll (1974) showed better discrim-

inability consistent with that obtained with older children. The reasons for this are not clear, although it is possible that the test is most useful when a brain insult occurs well after the development of cerebral dominance.

The relatively good lateralization of lesions by the dynamometer was somewhat surprising. Admittedly, the criterion of 1 standard deviation above or below the control mean is arbitrary. Furthermore, it leads to findings that are, if anything, conservative in implicating one cerebral hemisphere or the other. For example, the performance by the control group with the left hand on the Tapping Test was approximately .93 that of the performance on the right hand. The standard deviation was .09, so that any score from .84 through 1.02 was considered within normal limits, whereas scores less than .84 implicated the right cerebral hemisphere and scores greater than 1.02 implicated the left cerebral hemisphere. If one assumes that a person's right (preferred) hand averages 50 on the Tapping Test, an identical performance by the left hand would fall in the range of normal limits, whereas clinical interpretation would definitely suggest that the right hand was slow. Furthermore, the score with the left hand would have to be 41 or less in order to implicate the right cerebral hemisphere, whereas in clinical practice scores of 42 or 43 would certainly raise the question of slowness with respect to the left hand. With the procedure being somewhat conservative, it is not surprising to discover that 42% to 58% of the people evaluated in the lateralization analysis (Table 2) had performances that implicated neither cerebral hemisphere on each test. However, the conservative and arbitrary nature of the procedure would not appear to favor any particular test and is not likely to account for the fact that when decisions were made, they were correct in 86% of the cases for the dynamometer and in 65% and 76% of the cases for the Tapping Test and the Tactual Performance Test, respectively.

The question can be raised as to whether or not a cutoff score should be established for the dynamometer in the same fashion that it has been established for the Halstead measures. This appears unwise, because (a) the

number of subjects in the present study is too small to constitute an adequate standardization sample, (b) there are obvious sex differences that would require separate norms, (c) certain vocational and avocational activities of individuals may affect scores on this test, and (d) the accuracy of the Smedley dynamometers depends on a spring that may become weakened with use and lead to error in measurement. Therefore, no effort has been made to establish a cutoff score.

Overall, the hand dynamometer both discriminates between normal and brain-damaged persons and lateralized lesions as well as do existing measures. It appears to be a promising neuropsychological measure that warrants both clinical use and further formal evaluation, especially in consideration of the brief administration time required vis-à-vis the other two neuropsychological measures (Tactual Performance Test, Tapping Test) conventionally used in the Halstead-Reitan battery.

References

- Boll, T. J. Behavioral correlates of cerebral damage in children aged 9 through 14. In R. M. Reitan & L. A. Davison (Eds.), *Clinical neuropsychology: Current status and applications*. Washington, D.C.: V. H. Winston, 1974.
- Halstead, W. C. *Brain and intelligence: A quantitative study of the frontal lobes*. Chicago: University of Chicago Press, 1947.
- Reitan, R. M. An investigation of the validity of Halstead's measures of biological intelligence. *Archives of Neurology and Psychiatry*, 1955, 73, 28-35.
- Reitan, R. M. A research program on the psychological effects of brain lesions in human beings. In N. R. Ellis (Ed.), *International review of research in mental retardation* (Vol. 1). New York: Academic Press, 1966.
- Reitan, R. M. Psychological effects of cerebral lesions in children of early school age. In R. M. Reitan & L. A. Davison (Eds.), *Clinical neuropsychology: Current status and applications*. Washington, D.C.: V. H. Winston, 1974.
- Reitan, R. M., & Davison, L. A. (Eds.). *Clinical neuropsychology: Current status and applications*. Washington, D.C.: V. H. Winston, 1974.
- Vega, A., & Parsons, O. A. Cross-validation of the Halstead-Reitan tests for brain damage. *Journal of Consulting Psychology*, 1967, 31, 619-625.
- Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Received December 19, 1977 ■

An Initial Look at the Redundancy of Specialized MMPI Scales

James R. Clopton and Gary L. Klein
Texas Tech University

Numerous specialized Minnesota Multiphasic Personality Inventory (MMPI) scales have been developed despite speculation that the information provided by specialized scales replicates information that can be obtained from the 13 MMPI scales of the standard profile. In this study, scores of three specialized MMPI scales (the Prejudice scale, the Ego-Strength scale, and the MacAndrew Alcoholism scale) were found to be highly related to the scores of the standard MMPI scales. However, individual scores of the three specialized scales could not be accurately predicted from the standard scales. Furthermore, alcoholic and nonalcoholic psychiatric patients were more accurately identified by the 13 standard scales than by the MacAndrew Alcoholism scale.

The Minnesota Multiphasic Personality Inventory (MMPI) originally included three validity scales that assess test-taking attitudes and 10 clinical scales that identify common types of abnormal behavior. In addition to these 13 standard scales, numerous specialized MMPI scales have been constructed. Some of these specialized scales have been constructed to measure common personality dimensions such as dependency and prejudice. Other specialized MMPI scales have been developed to identify patterns of abnormal behavior, such as alcoholism, that are not assessed directly by any of the standard clinical scales. Interest in specialized scales has been increasing recently (Graham, 1978), and a number of specialized scales are routinely scored by automated MMPI interpretive systems.

The latest edition of the *MMPI Handbook* (Dahlstrom, Welsh, & Dahlstrom, 1975) listed 455 specialized scales that measure per-

sonality variables or identify patterns of abnormal behavior. Many of the specialized scales are too limited for widespread use, but the use of some specialized MMPI scales is quite common. Three specialized MMPI scales in common use are the Prejudice scale (Gough, 1956), the Ego-Strength scale (Baron, 1953), and the MacAndrew Alcoholism scale (MacAndrew, 1965). The Prejudice scale was designed originally to measure anti-Semitic prejudice, but it also appears to assess the broader trait of rigidity in thinking. The Ego-Strength scale is frequently used to help predict the likelihood that a person will profit from receiving psychotherapy. The MacAndrew scale has been shown to differentiate between alcoholic and nonalcoholic patients in a variety of treatment settings.

Despite the attractiveness of developing new MMPI scales for specialized tasks and the popularity of some specialized MMPI scales, there are serious questions regarding the use of specialized MMPI scales. One basic question concerns the possible redundancy or superfluity of the information provided by the specialized scales. Caldwell (Note 1) asserted that he could predict many specialized scale scores so well from the scores of the 13 standard MMPI scales that the specialized scales did not appear to provide any information beyond that which

This study was supported in part by a Summer Research Assistantship awarded to the second author by the Texas Tech University Graduate School.

The authors thank Roger L. Greene, Donald H. Baucom, and Don Beal for their helpful suggestions.

Requests for reprints should be sent to James R. Clopton, Department of Psychology, Texas Tech University, Lubbock, Texas 79409.

Table 1
Comparison of Observed and Predicted Raw Scores

Scale	Sex	<i>n</i>	<i>M</i>	<i>SD</i>	Range	<i>R</i> ^a	<i>F</i>	Accuracy
Prejudice								
Observed	Male	170	8.52	4.41	0-22	.71	28.87*	
Predicted	Male	170	8.52	3.70	1.54-19.16			45.3
Ego-Strength								
Observed	Male	112	42.70	7.25	21-57	.66	14.64*	
Predicted	Male	112	42.70	5.89	26.75-52.51			53.6
Observed	Female	85	38.95	7.30	23-57	.67	10.87*	
Predicted	Female	85	38.95	5.96	25.62-49.70			68.2
MacAndrew								
Observed	Male	112	24.78	5.02	14-38	.62	12.29*	
Predicted	Male	112	24.78	3.95	16.01-35.04			59.8
Observed	Female	85	21.84	4.70	8-32	.46	4.72*	
Predicted	Female	85	21.84	3.20	14.47-31.83			60.0

Note. The accuracy of the predicted scores is the percentage of subjects whose observed scale score was within the 95% confidence limits of the value predicted from the regression equation.

* $p < .0001$.

could be provided by the standard scales. Unfortunately, Caldwell's data have not been published, and other data regarding the possible redundancy of specialized MMPI scales are lacking.

Method

MMPI data were obtained from 197 psychiatric patients and from 170 male applicants for jobs with police and fire departments. The psychiatric patients and job applicants have been evaluated at a regional mental health center during the last 7 years. MMPI data from the psychiatric patients were used to investigate the clinically oriented Ego-Strength and MacAndrew scales. MMPI data from the job applicants seemed more appropriate for an investigation of the nonclinical Prejudice scale.

Regression analysis was used to determine how well subjects' specialized scale scores could have been predicted from their scores on the 13 standard MMPI scales. For the Ego-Strength scale and the MacAndrew scale, separate analyses were performed for male ($n = 112$) and female ($n = 85$) psychiatric patients.

A review of the psychiatric patients' records revealed that 48 patients (24.4%) had difficulties directly related to alcohol abuse. Discriminant analysis was used to determine whether alcoholic and nonalcoholic patients are more accurately identified by the MacAndrew scale or by the 13 standard MMPI scales.

Results

Table 1 presents the results of the regression analyses. The observed scores for all

three specialized scales were significantly related to the scores of the 13 standard scales, and a large portion of the variance in each of the specialized scales was accounted for by the 13 standard scale scores.

The regression equations were used to derive predicted specialized scale scores for subjects. It was then determined, for each of the three specialized scales, whether each observed scale score was within the 95% confidence limits of the predicted score. Approximately 60% of the predicted scores for the Ego-Strength and MacAndrew scales were found to be accurate by this criterion (see Table 1). For the Prejudice scale, 45.3% of the predicted scores were accurate. Thus, many of the individual scores of the three specialized scales could not be accurately predicted from the standard scale scores.

A ready explanation for the inaccurate predictions of the specialized scale scores is that the variance among the observed scores of each specialized scale was underestimated by the predicted scores (see Table 1). Imperfect multiple correlation ($1.00 > R > -1.00$) and use of a regression equation with a least-squares prediction rule assured the reduced variance among the predicted scores (Hays, 1963, pp. 500-501). As a consequence of the underestimation of the variance in specialized scale scores, prediction was most accurate

for scores close to the mean scale score, and the prediction became increasingly less accurate as the scores departed from the mean.

The discriminant analysis with the 13 standard MMPI scales as predictors correctly classified all of the female psychiatric patients as alcoholic or nonalcoholic and correctly classified 91.1% of the male patients (85.3% of the male alcoholics and 93.6% of the male nonalcoholic patients). In contrast, the discriminant analysis using the MacAndrew scale as the predictor correctly classified most of the nonalcoholic patients (100% of the female nonalcoholic patients and 89.7% of the male nonalcoholic patients), but it identified few of the alcoholic patients (7.1% of the female alcoholic patients and 35.3% of the male alcoholic patients). The optimal cutoff score for the MacAndrew scale (the scale score that would most correctly classify patients) was 27 for female patients and 25 for male patients.

Discussion

In this study an attempt was made to predict the scores of three specialized MMPI scales from the scores of the 13 standard MMPI scales. The three specialized MMPI scales were found to be highly related to the 13 MMPI scales of the standard profile, but individual scores on the three specialized scales could not be accurately predicted from standard scale scores. The three specialized scales examined in this study appear to provide information not available from the standard MMPI scales.

Extreme scores on the MMPI scales are often the scores of most importance in a clinical setting. In this study the prediction of specialized MMPI scale scores became increasingly less accurate as the scores departed from the mean. However, this inaccurate prediction of extreme specialized scale scores should not be interpreted as an indication that predictions from specialized scales are superior to predictions from the standard scales. The more extreme the scale score, the larger the error of measurement it probably contains. From this perspective, the most extreme specialized scale scores are also the

least reliable, and, consequently, it is unrealistic to expect highly accurate predictions of extreme specialized scale scores.

In this study, the 13 standard scales were more accurate than the MacAndrew scale in identifying psychiatric patients who abused alcohol. Future research should seek to determine, for other specialized scales, whether specialized scale scores or the standard scale scores are more closely related to external criteria (independent measures of relevant personality variables or patterns of abnormal behavior). For example, a comparison could be made of the predictions of successful outcome in psychotherapy by the Ego-Strength scale and by the 13 scales of the standard MMPI profile. It is not possible to anticipate the outcome of such comparisons from the finding that specialized scale scores cannot be accurately predicted from the standard MMPI scales.

Reference Note

1. Caldwell, A. B. *Recent advances in automated interpretation of the MMPI*. Paper presented at the Fifth Annual Symposium on Recent Developments in the Use of the MMPI, Mexico City, Mexico, February 1970.

References

- Barron, F. An ego-strength scale which predicts response to psychotherapy. *Journal of Consulting Psychology*, 1953, 17, 327-333.
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. *An MMPI handbook: Research applications* (Vol. 2, rev. ed.). Minneapolis: University of Minnesota Press, 1975.
- Gough, H. G. A personality scale for anti-Semitic prejudice. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine*. Minneapolis: University of Minnesota Press, 1956.
- Graham, J. R. Review of Minnesota Multiphasic Personality Inventory special scales. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 4). San Francisco: Jossey-Bass, 1978.
- Hays, W. L. *Statistics*. New York: Holt, Rinehart & Winston, 1963.
- MacAndrew, C. The differentiation of male alcoholic outpatients from nonalcoholic psychiatric outpatients by means of the MMPI. *Quarterly Journal of Studies on Alcohol*, 1965, 26, 238-246.

Received December 20, 1977 ■

Effects of a Self-Control Manual, Rapid Smoking, and Amount of Therapist Contact on Smoking Reduction

Russell E. Glasgow
North Dakota State University

This study evaluated a self-help treatment manual consisting of stimulus control, rapid smoking, and coping relaxation techniques. Sixty-nine subjects who smoked at least 20 cigarettes per day were randomly assigned to (a) a self-help manual with minimal (two sessions) therapist contact, (b) a self-help manual with high (seven sessions) therapist contact, (c) a high-therapist-contact rapid smoking condition, or to (d) a high-therapist-contact normal-paced smoking condition. Results indicate that while the overall program was moderately effective, groups did not differ on percentage of baseline smoking or on number of subjects abstinent at posttreatment, 3-month, or 6-month follow-up. Informant reports of subjects' smoking behavior and carbon monoxide analyses of expired air samples confirmed these findings. Subjects in the minimal contact condition generally followed through on their programs, required less therapist time, and were at least as successful as those in other groups in terms of long-term results. The implications of these findings for self-help manuals for smoking reduction are discussed.

Since the Surgeon General's report on the health effects of cigarette smoking in 1963, there have been numerous evaluations of smoking reduction programs. Reviews of this literature (Bernstein, 1969; Lichtenstein & Keutzer, 1971; Schwartz, 1969) have concluded that with few exceptions, the long-term effects of these studies have been disappointing. The typical result has been short-term reduction to 10%-40% of baseline, with relapse to approximately 75% of baseline at a 4- to 6-month follow-up (McFall & Hammen, 1971). A more recent evaluation

(Hunt & Belpaer, 1974) found that initial decreases in smoking frequency generally dissipate rapidly and asymptote around 60% of baseline rate. These authors also noted that on the average only 20%-30% of those subjects abstinent at termination were still not smoking at follow-up. In light of these findings, recent reviews (Bernstein & Glasgow, in press; Bernstein & McAlister, 1976; Lichtenstein & Danaher, 1976; Schwartz, 1977) have suggested that investigators focus on ways to maintain initial treatment effects.

One attempted solution to the maintenance problem has involved the use of multicomponent self-help treatment manuals. There are several potential advantages to such programs. By playing a more active role in a program, clients may learn skills and treatment techniques more thoroughly. With self-help programs there is a less abrupt transition period when treatment ends than with therapist-administered programs, and clients have their manuals to refer to if maintenance problems arise. It has also been hypothesized that client-directed treatments might lead to greater maintenance of therapeutic gains than

This report is based on a doctoral dissertation submitted to the Department of Psychology, University of Oregon. Appreciation is expressed to committee members David Campbell, Gerald Rosen, Herbert Severson, and especially Edward Lichtenstein, committee chairman. Nicole Barde, Sharon Brazil, Chris Clarke, Suzanne Dieringer, Bruce Fitzgerald, Gerald Gay, Cathy Hanauska, and John Piel served ably as therapists and research assistants. Thanks are also due Gerald Boyem and the Lane County Regional Air Pollution Authority for their kind analysis of the expired air samples.

Requests for reprints should be sent to Russell E. Glasgow, Department of Psychology, North Dakota State University, Fargo, North Dakota 58105.

therapist-directed programs by producing more internal attributions of success (Kopel & Arkowitz, 1975). Although a number of studies have been conducted in this area (Conway, 1977; Danaher, 1977; Harris & Rothburg, 1972; Ober, 1968; Winet, 1973; Brengelman, Note 1; Conway & Morton, Note 2; Danaher & Lichtenstein, Note 3; Pechacek, Note 4), self-help manuals have generally not been found effective (Glasgow & Rosen, 1978). A possible explanation for this situation is that investigators may have sacrificed quality for quantity. Most manuals contain a smorgasbord of procedures including self-monitoring and hierarchical reduction; stimulus control suggestions; self-reward, self-punishment, and behavioral contracting strategies; aversive smoking techniques; thought stopping or other cognitive interventions; information on the hazards of smoking and reasons for not smoking; suggestions for alternative behaviors incompatible with smoking, and so forth. Subjects are generally not given explicit instructions on how to implement these strategies. Consequently, subjects may not learn any of the techniques well enough to use them effectively once treatment has terminated.

The main purpose of this investigation was to evaluate the efficacy of a self-help smoking reduction manual that presented in depth a few promising techniques in an organized, sequential manner. The manual included stimulus control, rapid smoking, and coping relaxation techniques.

Stimulus control procedures have been found effective in reducing smoking until one reaches the level of 10–12 cigarettes per day (Marston & McFall, 1971; Sachs, Bean, & Morrow, 1970). In the present study stimulus control techniques were used as an initial component to assist in reduction of smoking levels and to teach subjects a behavioral problem-solving strategy. Rapid smoking has generally been found to be the single most effective cessation technique developed to date (Danaher, *in press*; Lichtenstein, Harris, Birchler, Wahl, & Schmahl, 1973; Schmahl, Lichtenstein, & Harris, 1972). It was included as a way to produce initial cessation and to get subjects to devalue the act of smoking. The relaxation procedure was in-

cluded as a coping skills technique (Goldfried & Trier, 1974) to provide an alternative response to smoking when subjects experienced a craving for a cigarette.

Another potential advantage of self-help treatment manuals is substantial savings in terms of therapist time. Survey data have indicated that many persons wishing to stop smoking would not attend a smoking clinic but would use a manual (McAlister, 1975). Thus, many more clients than are currently seen by therapists could be treated at a greatly reduced cost. Although several smoking manuals exist (Glasgow & Rosen, 1978), few of them have been tested under self-administered or minimal therapist contact conditions. Brengelman (Note 1) has tested a manual that was mailed to participants. Initial reports of abstinence were impressive, but the manual did not appear to have been compared to therapist-administered treatments or control conditions. Danaher and Lichtenstein (Note 3) reported less encouraging results for coverant control treatment manuals administered under minimal contact conditions.

The self-control manual was tested under two levels of therapist contact. Therapist-administered rapid smoking and a normal-paced smoking condition were also included as "component control groups" against which to evaluate the manual, thus yielding four treatment groups. Self-monitored frequency of smoking was recorded throughout treatment and at a 3-month follow-up. A 6-month phone call follow-up provided a further estimate of long-term effects. Carbon monoxide analyses of expired air samples and informant reports of smoking provided additional indices of smoking behavior. Process measures included subjects' therapeutic expectancies, amount of therapist contact time, indices of how well subjects carried out assignments (follow through), and ratings of the unpleasantness of aversive smoking sessions.

Method

Subjects

Smokers were solicited through local media announcements of a smoking reduction program. Selection criteria included smoking at least a pack a

day; informed consent; and payment of a \$20 deposit, \$15 of which was refundable on completion of the project. Applicants with a history of cardiovascular or chronic respiratory problems (8.9% of those contacting the clinic) were excluded. Subjects meeting the above screening criteria were required to obtain their physician's consent before beginning the program. Of those seeking physician approval, 8.1% were denied, most often due to pregnancy.

Sixty-two smokers, 32 men and 30 women, comprised the final sample. They averaged 32.6 years of age, estimated their baseline smoking rate at 31.0 cigarettes per day, had a self-monitored baseline of 24.7 cigarettes per day, and had smoked for an average of 15.1 years.

Therapists

Therapists were three male and three female undergraduate psychology majors. Therapists received extensive training over a 2-month period prior to the study. Training involved reading relevant background material, observing demonstrations of procedures, using role-playing techniques in the various groups, and seeing at least one pilot case. Treatment was standardized across therapists by means of procedural outlines for each session. Adherence to therapeutic procedures was further ensured by weekly group meetings and individual feedback from the project supervisor, who observed sessions or listened to audiotapes of a majority of the sessions. Therapists saw approximately an equal number of male and female subjects in each treatment condition.

Procedure

Subjects were given a general description of the program and were informed of the selection criteria over the telephone. Interested subjects were then randomly assigned to one of four treatment groups. Following an individual intake session, subjects began monitoring their cigarette consumption. After a 1-week baseline, subjects met individually with their assigned therapist for their respective 3-week treatment programs.

Treatment Groups

Minimal contact self-control ($n = 15$). This group received a 37-page manual detailing a multi-component treatment program for nonsmoking (Glasgow, Lichtenstein, & Danaher, Note 5). The manual presented a behavioral analysis of smoking and emphasized the importance of recording and counteracting smoking urges. Initial chapters focused on training in progressive relaxation and stimulus control techniques for hierarchical reduction. A three-phase relaxation training program presented relaxation as a coping strategy for use when experiencing urges to smoke.

The rapid smoking procedure closely followed that described by Kopel (1975). One trial consisted of

puffing on a cigarette every 6 sec until the subject could not bear to continue or until 10 min had elapsed, whichever came first. Immediately after each trial subjects completed a checklist of possible negative sensations (Glasgow, 1977). After a 2- to 4-min rest period, subjects underwent a second trial identical to the first, with the exception of a 5-min time limit. There were six such sessions consisting of two trials each. Sessions were held on the 1st, 2nd, 4th, 7th, and 10th days after the initial session.

Subjects' progress through the manual and completion of relaxation and rapid smoking sessions were recorded on a progress schedule contained in the manual. Subjects in the minimal contact group initially met with a therapist to receive their manual, a rationale for the program, and a demonstration of relaxation procedures. They then worked on their own, meeting once more with their therapist midway through the program to receive their first rapid smoking session. Subsequent rapid smoking and relaxation sessions were self-administered by clients at home. Therapists called weekly to check on subjects' progress and to answer questions.

High contact self-control ($n = 15$). This group received the same manual as the minimal contact group but had regular meetings with a therapist. Subjects were assigned to read a section of the manual and then met with their therapist to implement the assignments in that section. Seven meetings were held over the 3-week treatment period. Subjects received more direction from therapists on relaxation and stimulus control procedures than did minimal contact subjects, but rapid smoking and relaxation sessions were held at home after initial demonstrations. Treatment techniques and the sequence of components were identical to those of the minimal contact group.

High contact rapid smoking ($n = 16$). This group was intended as a replication of clinic-administered rapid smoking as used in previous research. The procedure and spacing of rapid smoking were identical to that for the manual groups, but all sessions were therapist administered. There was a 9-day "preparation period" after an initial meeting for subjects in this group before beginning rapid smoking. This was to insure that all groups completed treatment at the same time. Flaxman (1978) found that such a waiting period before a "target date" for beginning rapid smoking was more effective than beginning rapid smoking immediately.

High contact normal-paced smoking ($n = 16$). This group received an "aversive smoking" procedure used in previous research (Danaher, 1977; Kopel, 1975; Lichtenstein et al., 1973) to control for non-specific treatment effects. It involved smoking at one's normal rate while focusing on the unpleasant aspects of the pure smoking experience. Subjects were instructed to smoke until they could not bear to continue or until 5 min had elapsed, whichever came first. Otherwise, the rationale, number and spacing of sessions, and procedures used were identical to those of the rapid smoking group. If subjects smoked faster than one puff every 15-20 sec, they were reminded to smoke at their normal rate.

Measures

Subjects' therapeutic expectancies were assessed after the first treatment session and midway through the program following the first rapid smoking or normal-paced aversive smoking session. Monitoring of number of cigarettes smoked continued throughout treatment and during the week after treatment ended. Subjects again monitored their smoking for a 1-week period 3 months after treatment had ended and were then scheduled for a final interview. At this meeting deposits were returned and unannounced breath samples were collected. Expired air samples, which were analyzed for carbon monoxide (CO) content, were collected to provide an objective index of recent smoking behavior (Danaher, Lichtenstein, & Sullivan, 1976; Lando, 1975; Kopel, Note 6). Details of the collection procedure are available in Glasgow (1978). Informant reports of smoking were obtained by mailing informants a one-page questionnaire that asked for estimates of subjects' recent smoking behavior. Information on smoking rates 6 months after treatment ended was obtained by telephone calls asking for subjects' estimates of their average daily cigarette consumption during the preceding week.

Results

Of the 69 subjects beginning the program, 62 completed treatment. Five subjects dropped out after the first session for reasons unrelated to treatment, 1 subject dropped out after the fourth session following her doctor's advice when she found out that she was pregnant, and 1 subject in the normal-paced group dropped out after four sessions

apparently because she was not helped by the program. Three-month follow-up information was obtained on 61 of the subjects.

Preliminary Analyses

One-way analyses of variance revealed no significant differences between treatment groups on demographic indices, smoking history variables, or baseline smoking rates. There were no between-groups differences on expectancy of success, either after the first treatment session or after the first aversive smoking session. All groups indicated moderate to extreme confidence in their ability to stop smoking in their program.

Measures taken from the negative sensations checklist completed during aversive sessions revealed significant between-groups differences (see Table 1) on average number of cigarettes smoked, number of negative sensations endorsed, and aversiveness ratings for both the first and second trials (averaged across sessions). Groups also differed on length of trial for the first trial only. Planned comparisons indicated that rapid smoking was significantly more intense (i.e., more cigarettes smoked, more sensations endorsed) than the normal-paced procedure on each of these measures. The normal-paced group smoked at an average rate of 1 puff every 24.2 sec, compared to the 6-sec pace set for rapid smoking groups.

Table 1

Treatment Group Means on Process Measures for Trials 1 and 2 Collapsed Across Sessions

Measure	Manual				Aversive smoking only			
	Minimal contact		High contact		Rapid paced		Normal paced	
	1	2	1	2	1	2	1	2
Expectancy of success								
After first treatment session ^a		5.7		5.1		5.8		5.6
After first aversive smoking session ^a		6.1		6.1		6.3		6.1
Length of trial in min. ^b	5.8	3.6	7.4	4.5	6.6	3.8	5.0	4.8
No. cigarettes smoked ^b	2.5	1.4	3.0	1.9	2.7	1.6	1.2	1.1
No. negative sensations checked ^{b,c}	10.2	11.5	12.2	13.3	8.6	9.9	6.4	7.9
Aversiveness rating ^b	5.9	6.4	5.9	6.1	5.7	6.3	4.8	5.4

^a Based on a 7-point rating scale where 1 = not at all, 4 = moderately, and 7 = extremely.

^b Significant between-groups differences were observed on this measure.

^c Possible total of 20.

Table 2
Means and Standard Deviations on Percentage of Baseline Smoking and Therapist Contact Time

Group	% baseline smoking				Therapist time ^a	
	Post		Follow-up		M	SD
	M	SD	M	SD		
Manual						
Minimal contact	19.6	32.1	46.8	50.8	92.7	32.7
High contact	19.7	39.0	48.7	45.5	188.1	33.3
Aversive smoking only						
Rapid	10.6	17.2	63.6	42.1	167.9	23.0
Normal	17.4	35.4	58.2	41.8	153.9	25.9

^a In minutes; between-groups differences significant at the .001 level.

Follow-through measures indicated that subjects in all groups completed most of their programs. Three measures of completion of homework assignments for subjects receiving the manual were obtained from entries in the progress sheet in the front of subject's manuals. Subjects averaged 93% completion of reading assignments, 95% completion of relaxation practice sessions, and 90% completion of aversive smoking sessions. There were no significant differences on follow-through measures between the minimal contact group and the high contact group. Subjects receiving clinic-administered rapid smoking and normal-paced smoking completed 100% and 94% of their aversive sessions, respectively.

Effectiveness and Efficiency of Treatments

Percentage of baseline smoking averaged across treatment groups was 16.7% (3.4 cigarettes per day) during the week after treatment, and 40% of subjects were completely abstinent. At the 3-month follow-up, subjects averaged 54.7% of baseline (12.4 cigarettes per day), and 28% were not smoking. Correlated *t* tests revealed a significant decrease in smoking rate across groups from baseline to posttreatment, $t(61) = 13.29$, $p < .001$. Even though subjects were still smoking significantly less at the 3-month follow-up than they had at baseline, $t(60) = 7.03$, $p < .001$, they were smoking significantly more than they had at posttreatment, $t(60) = 6.42$, $p < .001$.

Between-groups differences on number of cigarettes smoked, percentage of baseline smoking, and number of subjects abstinent failed to reach significance either at post-treatment or at the 3-month follow-up. The pattern of treatment means (see Table 2) suggested that rapid smoking alone was somewhat more effective than the other treatments at the end of the program, but the great variability in the data appears to have precluded between-group differences. By follow-up the situation had reversed itself, with the manual groups being somewhat superior to other conditions. Again, effects were much too variable to obtain significance. Analyses of CO concentrations and informant reports of smoking similarly failed to reveal significant between-groups differences.

It was possible to contact 50 of the 62 subjects for the 6-month phone contact. At this point in time, subjects reported averaging 70.4% of their baseline smoking rate. Only 16% of those contacted reported being abstinent. The trend for manual groups ($M_s = 66.9\%$ and 64% of baseline for minimal and high contact groups, respectively) to be slightly superior to control groups ($M_s = 80.1\%$ and 67.8% for rapid smoking and normal-paced smoking, respectively) continued, but it again failed to approach significance on any outcome measure.

Efficiency was indexed by computing amount of therapist time spent in contact with clients from progress reports completed by therapists immediately after each session. There was a highly significant effect on this

measure, $F(3, 58) = 30.4$, $p < .001$. Tukey post hoc tests (Winer, 1971) revealed that the minimal contact manual group was more efficient than all other treatments and that the normal-paced group required less therapist time than did the high contact manual group. The minimal contact group required an average of approximately $1\frac{1}{2}$ hours (including telephone calls) of therapist contact, slightly more than half the time required by other groups (see Table 2).

Relationships Between Measures

The Pearson product-moment correlation between CO concentrations and self-report of cigarettes smoked that day was .40 ($p < .05$). The correlation between subjects' report of time elapsed since they smoked their last cigarette and CO levels was $-.62$ ($p < .001$). A t test comparing abstinent and smoking subjects on CO levels was highly significant, $t(53) = 4.89$, $p < .001$. Mean CO levels were 29.9 parts per million (ppm) for smokers and 10.4 ppm for nonsmokers. Informant reports of subjects' smoking were obtained on 57 of the 62 subjects. Twenty-one subjects reported abstinence when their informants were contacted, and all of these reports were confirmed by informants.

In a search for predictors of treatment success, numerous demographic, smoking history, and process measures were correlated with outcome at posttreatment and follow-up. The only variable found to be consistently related to outcome was the expectancy rating immediately after the first aversive smoking session. This rating was modestly correlated with percentage of baseline smoking at posttreatment ($r = -.37$, $p < .01$) and at the 3-month follow-up ($r = -.26$, $p < .05$). Posttreatment smoking rates were moderately correlated with follow-up smoking rates ($r_s = .38$ and $.47$ for number of cigarettes and percentage of baseline, respectively).

Discussion

The overall magnitude of observed change was somewhat better than has typically been reported in the smoking literature (Hunt & Bespalec, 1974; McFall & Hammen, 1971)

but was less impressive than most previous investigations of rapid smoking (Lichtenstein et al., 1973; Harris & Lichtenstein, Note 7; Weinrobe & Lichtenstein, Note 8). There are several possible explanations for this result. The present investigation used undergraduate paraprofessional therapists, and it is possible that this led to diminished effectiveness. However, this is unlikely, as Weinrobe and Lichtenstein (Note 8) have successfully used paraprofessionals as rapid smoking therapists. Also, one of the more experienced therapists in the present study had previously served in studies that had produced more impressive absolute results. Analyses revealed that this therapist was no more successful than other therapists in the investigation.

A fixed number of rapid smoking sessions and trials per session were used rather than continuing until subjects reported abstinence as in earlier research. While offering greater experimental control, the standardized approach likely reduces the clinical impact of treatment. Other recent studies using rapid smoking with a fixed number of sessions have generally reported absolute results similar to those of this study (see Danaher, in press) in contrast to the more impressive earlier results. Investigations of the effects of varying termination criteria (e.g., Weinrobe & Lichtenstein, Note 8) deserve further attention.

The main factor militating against significant results in this investigation was the enormous within-group variability observed. This is a common finding in smoking reduction outcome studies (Bernstein & Glasgow, in press; Lichtenstein & Danaher, 1976) and suggests the need for identifying predictors of success for different treatments. Consistent with past research, the present study was relatively unsuccessful in identifying demographic or smoking history variables that correlate with the outcome of standardized programs. It is suggested that future research should pursue the alternative strategy of evaluating different methods of individualizing treatment programs. Investigators using the locus of control scale (Rotter, 1966) have been moderately successful in assigning

subjects to appropriate smoking reduction programs (Best, 1975). It is possible that a recent extension of the internal-external locus of control dimension to the specific area of health (Wallston, Wallston, Kaplan, & Maides, 1976) would be even more effective in tailoring treatments to subjects.

It may be that features in the design of the manual reduced its effectiveness. One possibility is that portions of the manual may have been too demanding or complicated. Even though the program was restricted to a few techniques, these were covered in detail, and subjects may not have been provided with sufficient time to learn them adequately. The process of identifying effective (and ineffective) techniques in multicomponent treatment programs remains a complicated but important task for future research.

This study is consistent with numerous other investigations in finding only short-term (Winet, 1973; Conway & Morton, Note 2) or no beneficial effects (Conway, 1977; Danaher, 1977; Ober, 1968; Danaher & Lichtenstein, Note 3) from manuals. It must be concluded that at the present time effective self-help manuals for smoking reduction do not exist. It is suggested that authors concentrate on evaluating alternative treatment approaches before rushing to publish unvalidated programs. It may be that there is an addictive component of smoking—likely associated with nicotine (Russell, 1974; Schacter et al., 1977)—that is resistant to self-control approaches. Possible alternative strategies would be to establish a goal of controlled smoking (Frederiksen & Petersen, Note 9) rather than complete cessation or the use of nicotine chewing gum as a supplement to smoking manuals.

Despite its failure to improve treatment success, the minimal contact self-help manual condition was very efficient and produced long-term results at least as good as those of clinic-based treatment conditions. Unlike many other studies on self-help manuals (Glasgow & Rosen, 1978), minimal contact subjects completed the great majority of their programs. The high follow-through percentages in this program may be attributable

to the progress chart in the manual, which specified the schedule that subjects should adhere to in completing their programs. The results of the present study suggest that if an effective smoking reduction manual were developed, evaluation under minimal contact or self-administered conditions would be warranted.

The failure of rapid smoking to produce results superior to those of the control group is somewhat puzzling in light of previous investigations using similar groups (Danaher, 1977; Kopel, 1975; Lichtenstein et al., 1973). This result appears to be a combined effect of rapid smoking being less effective and the normal-paced procedure being more effective (at follow-up) than in other studies. Process measures indicated that rapid smoking was more intense and aversive than normal-paced smoking, thus suggesting that the treatments did differ procedurally. The relatively good long-term performance of what was intended as a control group, along with the failure of process measures taken during aversive sessions to correlate with outcome, suggests that the normal-paced procedure may be better construed as an alternative treatment than as a control group. Its use might be recommended in cases in which rapid smoking is contraindicated (Lichtenstein & Glasgow, 1977).

Reference Notes

1. Brengelman, J. C. *Experiments for the cessation of smoking*. Paper presented at the Third World Conference Against Smoking, New York, June 1975.
2. Conway, J. B., & Morton, V. *Rapid smoking and self-management procedures for smokers differing in locus of control*. Unpublished manuscript, University of Saskatchewan, 1975.
3. Danaher, B. G., & Lichtenstein, E. *An experimental analysis of covert control: Cueing and consequence*. Paper presented at the annual meeting of the Western Psychological Association, San Francisco, April 1974.
4. Pechacek, F. *Specialized treatments for high anxious smokers*. Paper presented at the annual meeting of the Association for the Advancement of Behavior Therapy, New York, December 1976.
5. Glasgow, R. E., Lichtenstein, E., & Danaher, B. G. *How to stop smoking: A self-directed approach*. Unpublished manuscript, University of Oregon, 1976.

6. Kopel, S. A. *Carbon monoxide monitoring: Clinical and research utility for smoking reduction*. Paper presented at the annual meeting of the Association for the Advancement of Behavior Therapy, San Francisco, December 1975.
7. Harris, D. E., & Lichtenstein, E. *The contribution of nonspecific social variables to a successful behavioral treatment of smoking*. Paper presented at the annual meeting of the Western Psychological Association, San Francisco, April 1971.
8. Weinrobe, P. A., & Lichtenstein, E. *The use of urges as termination criterion in a rapid smoking treatment program for habitual smokers*. Paper presented at the annual meeting of the Western Psychological Association, Sacramento, California, April 1975.
9. Frederiksen, L. W., & Petersen, G. L. *Controlled smoking: The case for a new treatment goal*. Paper presented at the meeting of the Association for the Advancement of Behavior Therapy, New York, December 1976.

References

- Bernstein, D. A. The modification of smoking behavior: An evaluative review. *Psychological Bulletin*, 1969, 71, 418-440.
- Bernstein, D. A., & Glasgow, R. E. The modification of smoking behavior. In O. F. Pomerleau & J. P. Brady (Eds.), *Behavioral medicine: Theory and practice*. Baltimore: Williams & Wilkins, in press.
- Bernstein, D. A., & McAlister, A. The modification of smoking behavior: Progress and problems. *Addictive Behaviors*, 1976, 1, 89-102.
- Best, J. A. Tailoring smoking withdrawal procedures to personality and motivational differences. *Journal of Consulting and Clinical Psychology*, 1975, 43, 1-8.
- Conway, J. B. Behavioral self-control of smoking through aversive conditioning and self-management. *Journal of Consulting and Clinical Psychology*, 1977, 45, 348-357.
- Danaher, B. G. Rapid smoking and self-control in the modification of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1068-1075.
- Danaher, B. G. Research on rapid smoking: Interim summary and recommendations. *Addictive Behaviors*, in press.
- Danaher, B. G., Lichtenstein, E., & Sullivan, J. M. Comparative effects of rapid and normal smoking on heart rate and carboxyhemoglobin. *Journal of Consulting and Clinical Psychology*, 1976, 44, 556-563.
- Flaxman, J. Quitting smoking now or later: Gradual, abrupt, immediate and delayed quitting. *Behavior Therapy*, 1978, 1, 260-270.
- Glasgow, R. E. *A self-control manual and amount of therapist contact in smoking reduction*. (Doctoral dissertation, University of Oregon, 1977). *Dissertation Abstracts International*, 1978, 38, 5018B. (University Microfilms No. 78-02,521)
- Glasgow, R. E., & Rosen, G. M. Behavioral bibliography: A review of self-help behavior therapy manuals. *Psychological Bulletin*, 1978, 85, 1-23.
- Goldfried, M. R., & Trier, C. S. Effectiveness of relaxation as an active coping skill. *Journal of Abnormal Psychology*, 1974, 83, 348-355.
- Harris, M. B., & Rothburg, C. A self-control approach to reducing smoking. *Psychological Reports*, 1972, 31, 165-166.
- Hunt, W. A., & Bessalec, D. A. An evaluation of current methods of modifying smoking behavior. *Journal of Clinical Psychology*, 1974, 30, 431-438.
- Kopel, S. A. The effects of self-control, booster sessions, and cognitive factors on the maintenance of smoking reduction. (Doctoral dissertation, University of Oregon, 1974). *Dissertation Abstracts International*, 1975, 35, 4182B-4183B. (University Microfilms No. 75-3895)
- Kopel, S. A., & Arkowitz, H. The role of attribution and self-perception in behavior change: Implications for behavior therapy. *Genetic Psychology Monographs*, 1975, 92, 175-212.
- Lando, H. A. An objective check upon self-reported smoking levels: A preliminary report. *Behavior Therapy*, 1975, 6, 547-549.
- Lichtenstein, E., & Danaher, B. G. Modification of smoking behavior: A critical analysis. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 3). New York: Academic Press, 1976.
- Lichtenstein, E., & Glasgow, R. E. Rapid smoking: Side effects and safeguards. *Journal of Consulting and Clinical Psychology*, 1977, 45, 815-821.
- Lichtenstein, E., & Keutzer, C. S. Modification of smoking behavior: A later look. In R. D. Rubin, H. Fensterheim, A. A. Lazarus, & C. M. Franks (Eds.), *Advances in behavior therapy*. New York: Academic Press, 1971.
- Lichtenstein, E., Harris, D. W., Birchler, G. R., Wahl, J. M., & Schmah, D. P. Comparison of rapid smoking, warm smoky air, and attention placebo in the modification of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1973, 40, 92-98.
- McAlister, A. Helping people quit smoking: Current progress. In A. J. Enelow & J. B. Henderson (Eds.), *Applying behavioral science to cardiovascular risk*. New York: American Heart Association, 1975.
- Marston, A. E., & McFall, R. M. Comparison of behavior modification approaches to smoking reduction. *Journal of Consulting and Clinical Psychology*, 1971, 36, 153-162.
- McFall, R. M., & Hammen, C. L. Motivation, structure, and self-monitoring: The role of nonspecific factors in smoking reduction. *Journal of Consulting and Clinical Psychology*, 1971, 37, 80-86.
- Ober, D. C. Modification of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1968, 32, 543-549.
- Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80(1, Whole No. 609).

- Russell, M. A. H. Realistic goals for smoking and health: A case for safer smoking. *Lancet*, 1974, 1, 254-258.
- Sachs, L. B., Bean, H., & Morrow, J. E. Comparison of smoking treatments. *Behavior Therapy*, 1970, 1, 465-472.
- Schacter, S., et al. Studies of the interaction of psychological and pharmacological determinants of smoking. *Journal of Experimental Psychology: General*, 1977, 106, 3-40.
- Schmahl, D. P., Lichtenstein, E., & Harris, D. E. Successful treatment of habitual smokers with warm smoky air and rapid smoking. *Journal of Consulting and Clinical Psychology*, 1972, 38, 105-111.
- Schwartz, J. L. A critical review and evaluation of smoking control methods. *Public Health Reports*, 1969, 84, 489-506.
- Schwartz, J. L. Smoking cures: Ways to kick an unhealthy habit. In, *Research on smoking behavior*. Rockville, Md.: National Institute on Drug Abuse, 1977.
- Wallston, B. S., Wallston, K. A., Kaplan, G. D., & Maides, S. A. Development and validation of the health locus of control (HLC) scale. *Journal of Consulting and Clinical Psychology*, 1976, 44, 580-585.
- Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.
- Winet, R. A. Parameters of deposit contracts in the modification of smoking. *Psychological Record*, 1973, 23, 49-60.

Received December 22, 1977 ■

A Social-Behavioral Analysis of Skill Deficits in Delinquent and Nondelinquent Adolescent Boys

Barbara J. Freedman

Bureau of Program Resources
Wisconsin Division of Corrections, Madison

Lisa Rosenthal

University of Wisconsin—Madison

Clyde P. Donahoe, Jr.

University of Wisconsin—Madison

David G. Schlundt

University of Wisconsin—Madison

Richard M. McFall

University of Wisconsin—Madison

Delinquent behavior is conceptualized as a manifestation of situation-specific social-behavioral skill deficits. The research was in two phases. In Phase 1, a measure consisting of 44 behavioral role-playing and problem-solving items—the Adolescent Problems Inventory (API)—was empirically developed, along with an item-specific criterion-referenced raters' manual. The inventory was designed to identify strengths and weaknesses in the personal and interpersonal skills repertoires of adolescent boys. Phase 2 was concerned with the validation of the API. In an initial validation study, the API responses of institutionalized delinquent boys were rated as less competent than the responses of either of two nondelinquent groups of teenage boys ("good citizens" and "leaders") from a public high school. Analyses of the inventory's characteristics showed it to be reliable, to be composed of items with little or no cluster structure, and to have extraordinary discriminant power. A second validation study compared the API responses of two groups: institutionalized delinquent boys who had frequent behavioral problems within the institution and institutionalized delinquent boys who had few acting-out problems within the institution. The former group was judged to respond less skillfully. A third validation study replicated previous group differences between delinquents and carefully matched nondelinquents. The study also showed that the type of directions given ("What would you do?" vs. "What is the best thing to do?") and test format (free response vs. multiple choice) significantly affected performance. It is suggested that researchers using a social skills conceptualization of personality do more thorough assessment studies of behavior pathologies before embarking on the development of large-scale social skills training programs.

It has been suggested that some individuals behave maladaptively simply because they lack the requisite skills to do better (e.g., McFall, 1976). In recent years, this skill-deficit conception of deviance has been reflected in numerous experimental skill-training programs aimed at treating such clinical populations as nonassertive college students (McFall & Twentymen, 1973), shy males (Twentymen & McFall, 1975), alcoholics (Sobell & Sobell, 1973), psychiatric inpatients (Goldsmith & McFall, 1975), and

male adolescent delinquents (Sarason & Ganzer, 1971).

Unfortunately, nearly all skill-training studies to date have been treatment oriented; that is, they have been concerned either with evaluating the general therapeutic utility of skill-training programs or with assessing the specific contributions of various training components, such as instructions, modeling, rehearsal, or feedback. Meanwhile, many fundamental questions concerning the underlying assumptions, concepts, and methods of the

skill-training approach have been ignored. Some investigators, for example, have developed the content of their skill-training programs without first conducting a thorough and systematic analysis of the performance problems supposedly addressed by the programs. As a result, they have had no way of knowing whether their programs actually focused on the most relevant problem situations for their clients or whether the behaviors taught in the programs represented genuine solutions to these target problems. Furthermore, some investigators have offered skill training without first establishing that their clients actually were deficient in the particular skills being taught.

What is needed at this point is research aimed at developing a taxonomy of the particular problem situations and skill deficits most characteristic of particular clinical populations. In the absence of such basic research, it will be difficult to develop valid methods for assessing and classifying the skill deficits of individual clients, and it will be difficult to determine what new behaviors clients need to acquire in order to perform more competently. Clearly, basic taxonomic research is a prerequisite to further treatment-oriented research.

The present research was based on a social-skills conception of delinquency among adolescent boys. Specifically, it was hypothesized that boys who have gotten into trouble with the law (i.e., adjudicated delinquents) would show situation-specific skill deficits when their performance in selected tasks was compared to that of matched nondelinquent

boys. The research was conducted in two phases. The first was concerned with identifying problem situations facing today's teenagers that might differentiate between the performance skills of delinquent and nondelinquent boys and, concomitantly, with developing explicit situation-specific criteria for evaluating performance competence. The first phase culminated in the creation of a behavioral role-playing measure of social skills, the Adolescent Problems Inventory (API). The second phase was concerned with validating the API by empirically evaluating its ability to differentiate between delinquent and nondelinquent boys. By implication, this phase also represented an indirect test of the utility of the underlying social-skills conception of delinquency. Although the research did not involve any treatment efforts, one of its long-term aims was to provide a solid foundation on which to build future skill-training programs.

Development of the Adolescent Problems Inventory

The procedures for the first phase were adapted from Goldfried and D'Zurilla's (1969) guidelines for the behavioral analysis of social competence. There were five sequential steps: (a) situational analysis, (b) item development, (c) response enumeration, (d) response evaluation, and (e) construction of the inventory and rater's manual. The purpose, procedures, and products of each step were as follows:

Situational analysis. The first step involved the identification of problem situations that might be related to delinquency. Initially, a large pool of promising situations was gleaned from a variety of sources: the sociological and psychological literature on the etiology of delinquency; case files of institutionalized delinquents; structured interviews with nondelinquent boys (ages 15-17); interviews with correctional psychologists, social workers, youth counselors, and teachers; and an open-ended questionnaire, given to 22 institutionalized delinquent boys (ages 14-18, $M = 15.8$, $SD = 1.2$), that asked about the problems of today's teenagers. This initial pool of situations subse-

The authors gratefully acknowledge the Wisconsin Division of Corrections and the Madison Public Schools for their support of the research project. They also wish to extend their thanks to Judith Rosenberg, who served as experimenter and data analyst for the second validation study. Barbara Freedman developed the Adolescent Problems Inventory and conducted the first two validation studies as part of her doctoral research. Lisa Rosenthal carried out Study 3 as part of her master's research. Pete Donahoe and David Schlundt performed extensive subsequent analyses of the data and assisted in the preparation of the manuscript. Richard McFall supervised the overall project. Requests for reprints should be sent to Barbara J. Freedman, Bureau of Program Resources, Division of Corrections, Box 669, Madison, Wisconsin 53701.

quently was reduced to a sample of 51 general descriptions of common problem situations; this reduction was achieved by eliminating redundancies, by condensing similar situations into a single version, and by excluding situations that seemed unrelated to social skills.

The ultimate aim was to develop an inventory not a scale. If we had been developing a scale, we probably would have selected maximally similar or highly-related items from among a pool of items assumed to represent a common domain, factor, or attribute. Since we were developing an inventory, however, we tried to select maximally dissimilar, nonoverlapping items. In addition, we did not assume the existence of any underlying factors, domains, or general attributes. We assumed only that the items had two things in common: (a) They were descriptive of problem situations with which many teenage boys are familiar, and (b) the problem situations were ones that if mishandled could get a teenage boy into trouble—conceivably into legal trouble. Thus, the assumed relationship between inventory performance and delinquency was essentially one of risk. The more frequently a teenager handles problem situations competently, the less likely he will get into trouble and be judged a delinquent.

The more universal and difficult a particular situation, the more appropriate it was considered to be for the purposes of the present research. Therefore, each of the 51 situations was rated on two 4-point scales by a new sample of 22 institutionalized delinquent boys (ages 14–18, $M = 15.8$, $SD = 1.2$). The first scale assessed how “common” they felt the situation was for boys their age; the second scale assessed how “difficult” they thought the situation would be for them to handle. A composite index of these common and difficulty ratings was constructed, and the 51 situations were rank ordered on that index to determine their appropriateness for inclusion in the study. Based on this ranking, nine situations were eliminated for being too uncommon or too easy, leaving a final pool of 42 problem situations.

Item development. The purpose of this second step was to translate the 42 general problem descriptions into specific items suit-

able for use as stimuli in a behavioral role-playing test. Ninety potential test items were written in the format of narrative descriptions in which the scene, characters, history, and goals of each incident were presented. At the end of each item, adolescents were asked to indicate how they would respond if they found themselves facing such a problem. The particular question posed to respondents took one of the two forms, depending on the particular item: (a) “What would you say or do now?” or (b) “What can you do to go about resolving this problem?” The following are examples of the two types of items: ¹

You're visiting your aunt in another part of town, and you don't know any of the guys your age there. You're walking along her street, and some guy is walking toward you. He is about your size. As he is about to pass you, he deliberately bumps into you, and you nearly lose your balance. What do you say or do now?

It is Saturday morning, and you have nothing planned for the whole day. There's nothing to look forward to all day. You feel bored already, just thinking about it. You need some kicks. What can you do to go about solving this problem?

Some items posed isolated problems, whereas other items posed related or sequential problems. For instance, the example in which the respondent was bumped into by a stranger was followed by this related item:

Now what if he had done the same thing, bumped into you, and you nearly lost your balance, and this time he said, “Look where you're going, clumsy!” What do you say or do now?

Response enumeration. The next step was to obtain a sample of possible responses, representing a wide range of social competency, to each of the 90 items. Twenty-three subjects participated in this step. There were 12 institutionalized delinquent boys (M age = 16.2, $SD = 1.0$); 6 nondelinquent boys from public and parochial high schools (M age = 16.9, $SD = .9$); and 5 adults (4 men, 1 woman) with professional experience in working with delinquent boys.

¹ Copies of all 90 items are available on request.

The items were administered to subjects in individual sessions lasting between $1\frac{1}{2}$ and $2\frac{1}{2}$ hours. Adolescents were instructed to respond as they typically would in each problem situation; adults were asked to play the role of "expert advisors" and give what they considered to be the best response for a teenage boy to make in each situation. The examiner read each item aloud, the subject responded orally, and the response was recorded verbatim. In addition, any questions or difficulties were noted and used as guides in subsequent efforts to refine and clarify instructions and item wording.

Response evaluation. The purpose of this step was to evaluate the quality of the 23 responses per item obtained in the preceding step. Judges, working independently, were asked to rate the competence of these responses using whatever subjective criteria they felt were important. Thirteen adults (4 men, 9 women) volunteered to serve as judges; they were advanced undergraduate psychology majors, clinical psychology interns, and professional psychologists. Depending on the amount of time that they were able to volunteer, each judge rated between 22 and 90 items. For any particular item, the number of judges ranged from 8 to 13, with a median of 11. Each judge was given a 145-page packet containing the 90 items along with typed transcripts of 23 responses to each. Judges worked independently, at their own pace, at home. They were blind as to the purpose of the study and the origins of the responses that they were evaluating. To minimize possible order effects, the 23 responses to each item were arranged randomly; furthermore, 5 judges progressed from Item 1 to Item 90, 5 went from Item 90 to Item 1, and 3 worked from Item 45 outward in both directions.

All of the judges first classified each of the 23 responses to each item as "competent," "incompetent," or "neither competent nor incompetent."² Then they rated the *relative* competence of each of the 23 responses per item on a 50-point scale, ranging from 0 = *maximally incompetent* to 50 = *maximally competent*. In addition, four judges were asked to specify in writing, as explicitly as possible, which criteria they had used to

evaluate all of the responses to a particular item, and a fifth judge orally explained the criteria that he had developed. The 50-point ratings and the written criteria were used later in the construction of a rater's manual.

Based on judges' evaluations, the original list of 90 items was reduced to a final set of 44 items. Interjudge agreement and item difficulty were the two criteria used for item selection. An item was retained (a) if there was 75% or higher interjudge agreement in the competency classification of the 23 responses to it and (b) if 25% or more of the 23 responses to the item were judged to be incompetent by at least 75% of the judges.

Construction of the inventory and rater's manual. The 44 items meeting both criteria were rewritten and polished in a final effort to eliminate any ambiguities or other problems noted in preceding steps. These items, in their final form, comprised the API, which was designed to be administered as a behavioral role-playing test.

A rater's manual for the API was also developed in this step. The manual presented explicit criteria for rating subjects' responses to each of the 44 items on a 5-point scale. Scale values were 8, 6, 4, 2, and 0; these values corresponded to judgments ranging from very competent to very incompetent. The criteria, which were item specific, were based on data provided by judges in the preceding response evaluation step. The manual's criteria reflected the distinguishing characteristics of the responses that judges had identified as competent, incompetent, or neither. Finer distinctions among responses were made by taking into consideration judges' ratings on the 50-point scale. Whenever possible, the manual incorporated judges' statements concerning the criteria that they used in their evaluation of responses to particular items. Sample responses representing different levels of competency were also included in the manual.

² Judges were allowed to sort the responses without concern for their distribution across the three categories; for example, it was possible for a judge to classify all 23 responses to a particular item as "competent."

To illustrate the item format and scoring manual, one item (12) is presented in full below. In addition, Table 1 presents a summary abstract of all 44 problem situations retained in the final version of the API.³

It is 1:30 at night, and you're walking along a street near your home. You're on your way home from your friend's home, and you know it is after curfew in your town. You weren't doing anything wrong. You just lost track of time. You see a patrol car cruising along the street and you feel scared, because you know you can get into trouble for breaking curfew. Sure enough, the car stops next to you, the policeman gets out, and he says, "You there, put your hands on the car. Stand with your feet apart." What do you say or do now?

Score:

- 8—EITHER the subject does it without saying anything or he asks a brief general question respectfully.

EXAMPLE: "What's wrong, officer?" "Is something the matter?" or he explains honestly and convincingly where he was.

- 6—The subject explains where he was, etc., but in a less assertive or less convincing manner.

EXAMPLES "I just got out of Pete Jones' house. You can call him if you want to."

- 4—No specific criteria . . . midway between responses scored 6 and 2.

- 2—The subject is antagonistic or flippant or insolent.

- 0—EITHER the subject hits the policeman or he runs away.

Validation of the Adolescent Problems Inventory

In this phase of the research, the API was subjected to two tests of concurrent discriminant validity. The first was a study comparing the API responses of three groups of adolescent boys assumed to represent three points along a continuum of social competence: a group of institutionalized delinquents, a comparison group of nondelinquent peers ("good citizens"), and a group of nondelinquent adolescent "leaders." The second study involved a comparison between the API responses of two groups of institutionalized delinquents, who differed in terms of their frequency of their placement in a confinement cottage as a result of so-called acting-out behaviors, according to institutional records.

Study 1

Method

Subjects. There were three groups, with 20 Caucasian boys in each, for a total of 60 subjects.⁴ The three groups were as follows:

1. **Delinquents:** These boys were all residents of a state correctional institution for juvenile offenders. They ranged in age from 14.1 to 17.8 years ($M = 16.4$, $SD = .94$). According to Hollingshead's (Note 1) index of socioeconomic status, which is based on the head-of-household's educational level and occupation, the group's mean socioeconomic level was 50.9 ($SD = 10.52$); this indicates that the typical delinquent subject came from the working class (Class IV). The various offenses for which the boys had been institutionalized were car theft, sale and possession of drugs, burglary, robbery, battery, vandalism, truancy, runaway, forgery, and arson. It was the first stay in the state correctional institution for 14 of the boys, the second stay for 3, the third for 2, and the fourth for 1 ($M = 1.5$, $SD = .88$). The length of stay at the time of testing ranged from 3 days to 13 months.

2. **Good citizens:** These nondelinquent boys were attending a public school. They were selected by the school's guidance counselors, who were asked to nominate individuals who were law-abiding, mature, responsible, able to get along well with peers and adults, and involved either in extracurricular activities or after school jobs. Their ages ranged from 14.3 to 17.7 years ($M = 16.4$, $SD = .94$). The mean socioeconomic level for the group was 45.7 ($SD = 12.03$); like the delinquent subjects, they tended to come primarily from the working class. Prior to testing, each nondelinquent was asked whether he had ever experienced legal difficulties; no one admitted to having a police record.

3. **Leaders:** These subjects were also public high school students selected on the basis of guidance counselors' nominations. In addition to possessing all of the attributes of the good citizens, these subjects were recognized as student leaders; they were the editors of the school newspaper and yearbook, student senators, class presidents, and star athletes. Their ages ranged from 14.8 to 17.9 years ($M = 16.8$, $SD = .87$).

Preliminary analyses comparing the ages and socioeconomic levels of the three subject groups revealed only one significant difference: The mean socioeconomic status of the leaders ($M = 34.4$, $SD = 20.13$) was higher than that of the other two groups ($p <$

³ The Adolescent Problems Inventory was copyrighted by the first author and may not be used in any form without her written permission.

⁴ This initial study was limited to Caucasian boys due to the availability of subjects and to a concern about not examining too many variables in a single study. Obviously, the experimental results can be generalized only to Caucasian boys until subsequent research examines the relevance of the Adolescent Problems Inventory for non-Caucasian boys.

Table 1

Abstract of 44 Problem Situations Covered in the Final Version of the Adolescent Problems Inventory

1. A male, peer, stranger deliberately bumps into you on the street.^a
2. Same as #1, plus he blames you.^a
3. A gym teacher picks on you, makes you do extra pushups.^{a,b}
4. A friend suggests buying booze illegally.^a
5. Your father tells you to stay home on Saturday night.^a
6. You want to break up with your girlfriend without hurting her.^{a,b}
7. The school principal threatens to suspend you for hassling a substitute teacher.^a
8. You come home late at night and your father is waiting up for you and is angry.^{a,b}
9. Your are called names by some guy in the schoolyard.
10. Your mother tells you to put on decent clothes before leaving the house.^{a,b}
11. A friend wants you to deliver some drugs; he offers drugs and money in return.^a
12. You are stopped on the street by a policeman after curfew.^a
13. Your father wants you to stop seeing one of your male friends.^a
14. Another boy makes an insulting remark about your mother.^a
15. A friend suggests that you two steal a handgun from a discount store.^a
16. You back your car over the neighbor's trash can; he yells at you.^{a,b}
17. Your friend is upset because you dated a girl he likes.^a
18. You've been grounded. A friend urges you to sneak out of the house.^a
19. Your father gives you an ultimatum about getting your hair cut.^a
20. A policeman comes to your door and asks for you.^a
21. A teacher accuses you of writing obscene words on the walls in the men's room.^a
22. A friend suggests joy riding in a car with the keys left in it.^a
23. You run out of gas, get to work late, and get fired.^a
24. Your father gets upset when you ask to borrow the car.^a
25. A friend asks you to steal something for him from where you work.^a
26. While with a friend, your father angrily tells you to go clean your room.^a
27. An older friend asks you to help hold up a gas station.^a
28. You want to ask the manager of a McDonald's for a job.^a
29. Your girlfriend offers you a joint at a party.^a
30. You ask a girl for a date and she says that her father won't let her go out with you.^{a,b}
31. A girl's father meets you at the door and says he won't let her go out with you.^a
32. Peers at school hassle you about your criminal record.^a
33. A job interviewer is biased by your criminal record.^a
34. A teacher hassles you about your criminal record.
35. You wake up in a bad mood.^{a,b}
36. You need more money, your parents can't give it to you, and you are too young for a regular part-time job.^a
37. You are bored and want some fun.^a
38. You are studying for a final exam. A friend wants you to go to a concert instead.^a
39. Your mother forbids you to see a friend again.^a
40. Your girl breaks up with you. You feel miserable.^a
41. You don't feel like delivering your paper route today.^a
42. You feel hopelessly lost in a geometry class.^a
43. You have a car and want something exciting to do.^a
44. Your mother hassles you about going to church.^a

^a Good citizens > delinquents.

^b Leaders > good citizens.

.005); the leaders tended to come from the upper middle class (Class II).

Procedure. The API was administered to subjects by a female examiner in individual sessions lasting approximately 1 hour each.⁵ After presenting the instructions, the examiner played an audiotape containing the 44 test items. For items involving interactions with men, the tape-recorded voice was that of a man; for interactions involving women, the voice was a woman's. The examiner presented the test items

one at a time, using a remote control switch to start and stop the stimulus tape. Subjects' oral responses were recorded on a second machine for subsequent evaluation by trained "blind" judges.

⁵ It is an unanswered empirical question as to whether the use of an adult female examiner had any significant effect on the teenagers' responses. Future research must examine this question.

Working independently, two raters listened to subjects' recorded test responses and rated the competence of each response on a scale ranging from 0 (very incompetent) to 8 (very competent), according to the criteria outlined in the raters' manual. Intermediate rating values (1, 3, 5, and 7) were permitted if judges felt that responses were midway between categories. The raters were a male and a female, both seniors majoring in psychology. The protocols of 10 subjects (5 delinquents, 5 nondelinquents) were used for training. Interrater reliability for the remaining 50 protocols was analyzed by computing the Pearson product-moment correlation between raters' scores for individual responses to all 44 items across all 50 protocols (i.e., the correlation was among 2,200 pairs of ratings). There was a high level of interrater agreement ($r = .99$). The mean of the two raters' judgments was used in all subsequent analyses.

In addition, data on IQ and grade point average for the preceding year were obtained on most subjects from institutional and school records. Unfortunately, the IQ estimates for nondelinquent and delinquent subjects were based on different tests: For nondelinquents the estimates came from the Henmon-Nelson test; for delinquents the estimates came from the Culture Fair Test, a nonverbal measure, and from the Wide-Range Vocabulary Test, a verbal measure.

Results

Across all 44 items, delinquents earned a mean score of 2.73 ($SD = .77$); the mean for good citizens was 5.86 ($SD = .65$); and the mean for the leaders was 6.77 ($SD = .60$). Planned group comparisons, based on total API score, revealed that the leaders performed significantly better overall than did the good citizens, $F(1, 57) = 19.63$, $p < .001$. An item-by-item comparison revealed that the leaders significantly outperformed the good citizens on 7 out of the 44 individual items.

Good citizens, in turn, performed significantly better overall than did the delinquents, $F(1, 57) = 217.71$, $p < .001$. This latter comparison is the one of greatest interest, since these two groups were comparable in mean age and socioeconomic level. An item-by-item comparison showed that the good citizens also performed significantly better ($p < .05$) than the delinquents on 42 out of the 44 individual items. In absolute terms, the delinquent group's deficient performance was reflected in their earned mean rating of less than 4.0 (i.e., a value more indicative of incompetence than competence)

on 36 of 44 items. In contrast, good citizens earned mean ratings of less than 4.0 on only 1 item. Leaders did even better, earning no mean ratings below 4.0 and a mean of 5.0 or better on all but one item.

Table 1, which provides a brief abstract of the 44 API situations, also indicates which items yielded significant group differences ($p < .05$) in the comparisons between leaders and good citizens and between good citizens and delinquents.

Grade point averages and IQ scores for the three groups were obtained from institutional and school records. The mean IQ scores of the two nondelinquent groups were not significantly different: leaders = 111.8 ($SD = 6.91$) and good citizens = 102.7 ($SD = 16.93$). However, leaders earned significantly higher grade point averages than good citizens: 3.25 ($SD = .50$) and 2.75 ($SD = .63$), respectively; $F(1, 35) = 5.20$, $p < .05$. The good citizen group, in turn, had a significantly higher mean IQ than the delinquent group, whose mean score on the Culture Fair Test was 93.2 ($SD = 10.13$) and on the Wide-Range Vocabulary Test was 87.4 ($SD = 10.09$); $F_s(1, 35) = 4.33$ and 11.12, respectively (both p s $< .05$). The delinquent group also had a significantly lower grade point average ($M = 1.80$, $SD = .93$), $F(1, 35) = 25.83$, $p < .05$. Overall, there was a strong correlation between total API scores and verbal IQ scores ($r = .70$, $p < .05$).⁶ When correlations were computed separately within the delinquent and the nondelinquent samples, however, the significant relationship disappeared (for delinquents, $r = .14$; for nondelinquents, $r = -.03$).

The troublesome relationship between IQ score and API performance could not be ruled out as a contributing factor in the obtained group differences on the API. Nor could the effect of IQ be satisfactorily controlled through statistical manipulations, such as analysis of covariance, because of the reasons cited by Lord (1967). Despite these

⁶ In this correlation, nondelinquent IQ scores were based on the Henman-Nelson test; delinquent scores were based on the Wide Range Vocabulary Test. When delinquents' IQs were estimated by the Culture Fair Test, the correlation was somewhat lower (.58).

limitations, the IQ data were examined in another way for heuristic purposes. Two subsamples of eight subjects each were drawn from among the most intelligent of the delinquents and the least intelligent of the nondelinquents; in this way, two subsamples equated on mean IQ were composed (delinquents' $IQ = 101.6$; nondelinquents' $IQ = 101.0$). The mean API scores of these groups (2.51 and 6.60, respectively) were still highly discrepant, $F(1, 14) = 85.00$, $p < .05$. Of course, these results are only suggestive, due to the problem of statistical regression and the fact that the selected subsamples may not have been representative of their parent groups. These results tend to imply that a subject's verbal skills may contribute to his scores on both the IQ and API measures, but the API seems to measure something above and beyond verbal intelligence.

Characteristics of the Inventory

Reliability. The API was designed to be an inventory, rather than a scale; nevertheless, a reliability analysis was performed using the entire sample of 60 subjects. The results of the analysis appear in Table 2. These statistics should be interpreted with caution, since they were computed on a sample containing only extreme groups rather than on a random sample of adolescent boys. The effect of using extreme groups is to inflate estimates of internal consistency such as the coefficient α and the corrected item-total correlation.

Item relationships. Because the number of situations relative to the number of subjects was large, it was not appropriate to perform a factor analysis with the present data. However, four hierarchical cluster-analytic techniques were used in an attempt to group situations on the basis of similarity of subjects' performance competence. A complete linkage and an average linkage cluster analysis were performed on both the correlation and the squared Euclidian distance matrices (Lance & Williams, 1967; Sneath & Sokal, 1973). The situational clusters that emerged differed from one technique to the other and were generally uninterpretable in terms of their content. Between 8 and 14

Table 2
Reliability Analysis of the Adolescent Problems Inventory

Variable	<i>M</i>	<i>SD</i>
Total score	225.48	9.06
Item <i>M</i>	5.12	.75
Item <i>SD</i>	2.86	.48
Interitem <i>r</i>	.39	.15
Corrected item-total <i>r</i>	.62	.13

Note. Coefficient alpha = .966.

weak clusters were identified, depending on the solution used. For example, one of the clusters in the complete linkage cluster analysis of the correlation matrix consisted of four situations: (a) "Your father forbids you to go out with a friend," (b) "you're embarrassed to ask a teacher for help," (c) "you're fired by your boss for accidentally being late to work," and (d) "a policeman comes to your door and asks for you." The majority of clusters showed a similar lack of interpretability. Even the two individual items with the strongest association ($r = .79$) were not related in a readily apparent way: #19. Your father gives you an ultimatum about getting your hair cut; #29. Your girlfriend offers you a joint at a party.

The lack of clear results in the cluster analyses indicates that competence scores are not the proper measure on which to construct a situational taxonomy. Instead, one probably would do better to classify situations either on the basis of the similarity among their stimulus properties or on the basis of the similarity of the specific behavioral task requirements of the situations. The lack of consistent clustering of the situations, in conjunction with their moderate intercorrelations, leads to the conclusion that competence scores show a rather high degree of situational specificity, especially when the items are specifically designed to be nonoverlapping in their content, as in the API.

Discriminating power. Although planned comparisons indicated that good citizens performed significantly better than delinquents on 42 of the 44 API items, the significant F s do not reveal the degree to which the API can actually discriminate between these groups. A discriminant analysis, which is

more appropriate, was performed using four sets of items from the API. Mean scores were computed for each of the 20 delinquents and 20 good citizens across 6 situations involving aggression (AGR), across 21 situations involving interactions with adult authorities (AUT), across 2 situations involving bad moods, (BM), and across 8 situations involving resistance to temptation (RTT). The resulting discriminant function was $Y = -.10AGR - .32AUT - .02BM - .14RTT + 2.53$. Probabilities of misclassification for this function were computed using the \bar{U} method of Lachenbruch and Mickey (1968), who showed that this method is superior to the method of validating the function by resubstituting the original data. The \bar{U} method gives good estimates of the results that would be obtained by using a fresh validation sample.

Results indicated that the estimated probability of misclassifying a delinquent is .00, and the estimated probability of misclassifying a good citizen is .11. In other words, the discriminant function based on the four content scores of the API was 89% correct when it was used to classify the subjects in the derivation sample. In particular, its success rate was 100% in correctly identifying the delinquents.

Since the proportion of delinquents in the derivation sample was .5, it would be misleading to use these results to estimate the performance of the API if it were applied to a more typical population of adolescent boys. Therefore, an additional analysis was performed using various base rates, selection ratios, and costs of misclassification, all of which are important factors that must be considered when evaluating the ability of an instrument such as the API to discriminate between delinquents and nondelinquents.

Considering the difficulties in correctly identifying a population with a very low base rate (Meehl & Rosen, 1955), the API can be expected to perform remarkably well. To illustrate, consider a hypothetical situation: Suppose there is a population in which the base rate of adjudicated delinquents is known to be .03, the selection ratio is fixed at .08, and it would cost 25 times as much to mis-

classify a delinquent as to misclassify a nondelinquent. Under these conditions, the results of our discriminant analysis indicate that the API could be expected to perform as follows if used to classify 100 boys: We could expect to be correct 95 times out of 100 in identifying delinquents and nondelinquents. This is superior to what we could expect if we were to try to identify boys on a random basis, in which case we would be expected to be correct only 90 times out of 100. A selection ratio of .08 means that we would be interested in selecting 8 out of every 100 boys (e.g., candidates for a training program). Applying our discriminant rule, 3 of these 8 would be delinquent (i.e., valid positives), and 5 would be nondelinquents. If we had elected instead to select 8 boys randomly, we could expect virtually no delinquents to be selected. Finally, the validity coefficient of the API, in this context, is .60, which is the mathematically derived, theoretical upper limit to the validity that *any* instrument could attain.

Study 2

Subjects in the preceding validation study were presumed to represent three levels of performance along the full continuum of social competence. The purpose of this second validation study was to determine how well the API could differentiate between two groups of boys representing less extreme points on that continuum. Specifically, the study compared the API performances of two groups of institutionalized delinquent boys who were known to differ in their history of disruptive behaviors and rule violations within the institution.

Method

There were two groups, with 15 Caucasian delinquent boys in each, for a total of 30 subjects. All were residents of a state correctional facility at the time, and none had been subjects in the first study. One group was comprised of boys whose records indicated that they had a history of frequently engaging in disruptive behaviors within the institution; that is, they had spent more than 25 days ($M = 50.13$ days) during the preceding 6 months in the institution's security cottage as a result of running away, possessing drugs or contraband, assaulting peers or staff, or other serious misbehaviors. The second group

was comprised of boys who had spent less than 5 days ($M = 1.74$ days) in the security cottage during the same period.

The boys ranged in age from 14 to 17 years. The mean age of the high-disruptive group was 16.46 years ($SD = 2.82$); for the low-disruptive group, it was 16.40 ($SD = .75$). A comparison of the mean socioeconomic backgrounds of the two groups was not statistically significant; however, the high-disruptive boys tended to come from Class V families, whereas the low-disruptive boys tended to come from Class IV families. The testing procedure was identical to the one used in the first study. The two raters in the present study achieved absolute agreement on 93.5% of their ratings for individual responses to each of the 44 items by each of the 30 subjects.

Results

A one-way analysis of variance revealed that the low-disruptive subjects earned significantly higher total scores on the API than did high-disruptive subjects, $F(1, 28) = 5.92$, $p < .025$. Furthermore, in an item-by-item comparison, the low-disruptive group performed better on 32 of the 44 items, $\chi^2(1) = 8.64$, $p < .01$.

Study 3

This study addressed two important and unanswered questions from the previous studies. First, it assessed whether poor API performance is actually due to skill deficits, as hypothesized, or whether it is simply an artifact of the task's instructions. Standard API instructions are "Imagine that you're actually in the situation, and tell me in your exact words what you would say or what you would do if you were really there." Presumably, the delinquent subjects faithfully followed the instructions and reported their typical responses. It is possible that they would have given more competent responses, however, if they had been asked to give the *best* responses that they could think of, rather than their *typical* responses. This possibility was explored in this study.

Second, assuming that delinquents do have skill deficits, what is the specific nature of such deficits? In previous studies, the API was administered in an open-ended, free-response format. The present study examined whether delinquents would show comparable deficits if the API were administered in a

multiple-choice format. Delinquents may be deficient at *generating* competent responses, but would they also be deficient at *recognizing* competent responses?

The present study also introduced two methodological refinements to the basic paradigm used in the preceding studies. The first dealt with subject selection. In Study 1, good citizens were selected on the basis of guidance counselor nominations. Post hoc analyses indicated that the good citizens and delinquents were comparable on mean age and socioeconomic status but not on IQ. In the present study, each delinquent subject was carefully prematched on age, socioeconomic status, and IQ with a nondelinquent subject, who was drawn from public school files.

The second refinement was in the area of response ratings. In Studies 1 and 2, raters listened to the audiotaped responses of each subject to all 44 API items before proceeding to rate the next subject's responses. This procedure may have introduced a rating bias. For example, a subject's early responses may have contaminated ratings of his later responses. A subject's grammar, diction, or revelation of illegal behavior may have led a rater to give that subject a generally negative evaluation. To control such influences in the present study, typed transcripts of subjects' responses were rated. Furthermore, raters evaluated all subjects' responses to a given item before proceeding to the next item; the order of answers within each item was randomized.

Method

Subjects. The subjects were 40 delinquent boys from the Wisconsin School for Boys at Wales and 40 nondelinquent boys from the Madison Public School System. All were Caucasian, 16 or 17 years old, with IQs between 82 and 117. All were from families in which the head of the household was a small independent business person; clerical or sales worker; or skilled manual, semiskilled, or unskilled worker. Public school boys were asked if they had ever been in any trouble with the police; only those who indicated that they had not were included in the study.

Each delinquent subject was matched with a nondelinquent youngster of approximately the same age, IQ, and socioeconomic status. Of necessity, available scores were used to estimate subjects' IQs. Unfortunately, the intelligence tests previously administered to the two populations were not the same; the Wide-

Range Vocabulary Test was used for the delinquents, and the Lorge-Thorndike Intelligence Test was used for the nondelinquents. Socioeconomic status was estimated using the Occupational scale of the *Two-Factor Index of Social Position* (Hollingshead, Note 1).

Design. A $2 \times 2 \times 2 \times 2$ factorial design was used. The four factors were subjects (delinquent vs. nondelinquent); instructions (standard—"What would you do" vs. alternate—"What is best to do"); test format order (free-response format administered first vs. multiple-choice first); and item-group order. (The half of the items administered first to one group of subjects was administered second to the other group of subjects, and vice versa.) These four factors yielded 16 experimental cells, with five subjects per cell. Matched pairs of delinquent and nondelinquent subjects were randomly assigned to cells across the remaining three factors.

Each subject was administered all 44 API items under only one set of instructions—either standard or alternate. However, every subject was administered one half of the test in a free-response format and the other half in a multiple-choice format, with the format order balanced across subjects. To control for possible differences due to item content, the 44 API items were split into two equivalent groups of 22 items each. The presentation order of item groups (A and B) was balanced across subjects, instructions, and test formats.

Development of stimuli. The alternate instructions for the API told the subject to describe "What you think someone should say or do in that situation—not necessarily what you would do, but what you think is the best way someone could solve that problem?"

The multiple-choice version of the API was constructed by writing five response alternatives for each of the 44 items. The five options per item were derived from the response criteria given in the API scoring manual. The five responses corresponded to the points on the manual's 5-point rating scale: very competent, competent, neither competent nor incompetent, incompetent, and very incompetent.

Item Groups A and B, with 22 items each, were constructed as follows: All API items were classified according to the type of interaction involved (e.g., with parents, teachers, male peers, girls, police, etc.). The items from each category were divided as equally as possible between Item Groups A and B. The equivalency of the resulting groups was then checked by examining the scores on A and B items previously earned by subjects in Study 1: delinquents: $t(42) = .408$; good citizens: $t(42) = .973$; and leaders: $t(42) = .198$ (all *ns*). The item groups appeared to be satisfactorily equivalent.

Procedure. The procedure was identical to that in Study 1, except for the following important modifications or additions:

Each subject was individually administered one of the eight possible versions of the API by a female experimenter. To control for differences in reading ability, the subject listened to an audiotaped presentation of the API at the same time that he read from

the written test. If multiple-choice items were being presented, the five alternative responses also were presented on the audiotape, as well as in the written test.

During the free-response half of the test, the subject responded orally to each item, and his response was tape-recorded for subsequent transcription and scoring by raters. In the multiple-choice half of the test, the subject responded to each item by circling the letter corresponding to his response choice on the test booklet. The experimenter subsequently scored the multiple-choice responses with a key, assigning values on a 5-point scale (0, 2, 4, 6, and 8), corresponding to the competence criteria in the original API scoring manual.

Typed transcripts were made of subjects' oral responses under the free-response format. These transcripts then were edited to remove conspicuous use of grammar, word choices, or self-disclosure that might inadvertently reveal the identities of delinquent subjects. Finally, these typed responses were grouped by item and randomized for subsequent rating by four volunteers from an advanced undergraduate psychology class. One pair of raters rated half of the items; the other pair rated the rest. Across all items in Group A, the mean interrater agreement was .81, with the intercorrelations on individual items ranging from .64 to .93. Across all items in Group B, the mean agreement was .74, with interrater agreement on individual items ranging from .49 to .94. When raters were asked to repeat their ratings of selected items, as a way of estimating intrarater reliability, the mean was .90, with a range between .76 and .96.

Results

To evaluate whether the subject selection procedures yielded equivalent groups, t tests were performed on the delinquent and nondelinquent groups for the three matching variables of age, IQ, and socioeconomic status. There were no significant differences on age, $t(39) = .545$, or socioeconomic status, $t(39) = .206$. There was, however, a significant difference on IQ scores, $t(39) = 2.967$, $p < .05$, with nondelinquents scoring higher. Although the IQ difference between groups was statistically significant, the absolute magnitude of the difference was too small to be very meaningful. One standard deviation on the intelligence tests that were used is 15 points; yet, the mean IQ difference between delinquent subjects and their matched controls was only 3.8 points, with a range between 0 and 13 points. Moreover, a Pearson product-moment correlation computed across all subjects between IQ and total API scores

Table 3

Mean Response Ratings and Standard Deviations for Subjects in Each Testing Condition

Group	Test A		Test B		Test A		Test B	
	1	2	1	2	1	2	1	2
Free-response format								
Delinquents								
<i>M</i>	2.90	3.47	3.98	4.62	4.55	4.65	4.06	4.83
<i>SD</i>	.74	1.23	1.10	1.03	1.50	1.25	.94	1.42
Nondelinquents								
<i>M</i>	4.73	4.82	4.88	5.80	5.29	4.46	6.14	5.43
<i>SD</i>	.64	.40	.45	.58	.39	.54	.66	.44
Multiple-choice format								
Delinquents								
<i>M</i>	5.96	5.49	3.87	4.22	5.78	5.20	5.87	5.60
<i>SD</i>	.82	.53	1.38	2.22	1.30	.78	1.17	1.31
Nondelinquents								
<i>M</i>	6.38	5.42	6.09	5.96	6.29	6.27	5.91	6.16
<i>SD</i>	.49	.67	.55	.40	.60	.59	.81	.49

Note. 1 and 2 refer to items taken first and second, respectively.

was nonsignificant and relatively low (.06). This suggests that there was a negligible relationship between IQ scores and performance on the API.

Results were analyzed by two separate analyses of variance, one for free-response items and one for multiple-choice items; since it could not be assumed a priori that free-response and multiple-choice formats were equivalent, it was not appropriate to treat them as repeated measures within a single overall analysis of variance. Table 3 presents the mean response ratings and standard deviations for the subjects in each testing condition. The group means, collapsed across format order and item-group order, are graphically presented in Figure 1.

There was a significant performance difference between delinquents and nondelinquents both on the free-response version of the API, $F(1, 64) = 21.68$, $p < .001$, and on the multiple-choice version, $F(1, 64) = 10.59$; $p < .005$. Delinquents earned a mean rating of 4.13 on free-response items ($SD = 1.33$), and nondelinquents earned a mean of 5.19 ($SD = .76$). On the multiple-choice version, delinquent subjects had a mean score of 5.25 per item ($SD = 1.48$); for nondelinquents, the mean was 6.06 ($SD = .65$).

Within the free-response format, there was a significant main effect due to instructions, $F(1, 64) = 5.30$, $p < .05$, with the alternate instructions (which called for subjects to give the *best* response) leading to higher scores than the standard instructions (which asked subjects to tell how they would actually react in the various situations). This same instruction effect was not present in the multiple-choice version of the API, $F(1, 64) = 3.42$.

The performances of delinquents and nondelinquents also were compared under standard versus alternate instructions within each of the test formats (i.e., free response and multiple choice). The Kolmogorov-Smirnov two-sample test was used for these comparisons; it assesses the similarity of the distributions of scores for the two subject groups. Significant differences were found between delinquents and nondelinquents in the free-response format with both standard instructions ($p < .01$) and alternate instructions ($p < .05$). In the multiple-choice format, however, no significant differences were found for either set of instructions.

To compare within-group performance on the multiple-choice and free-response formats, matched t tests were performed. Delinquents

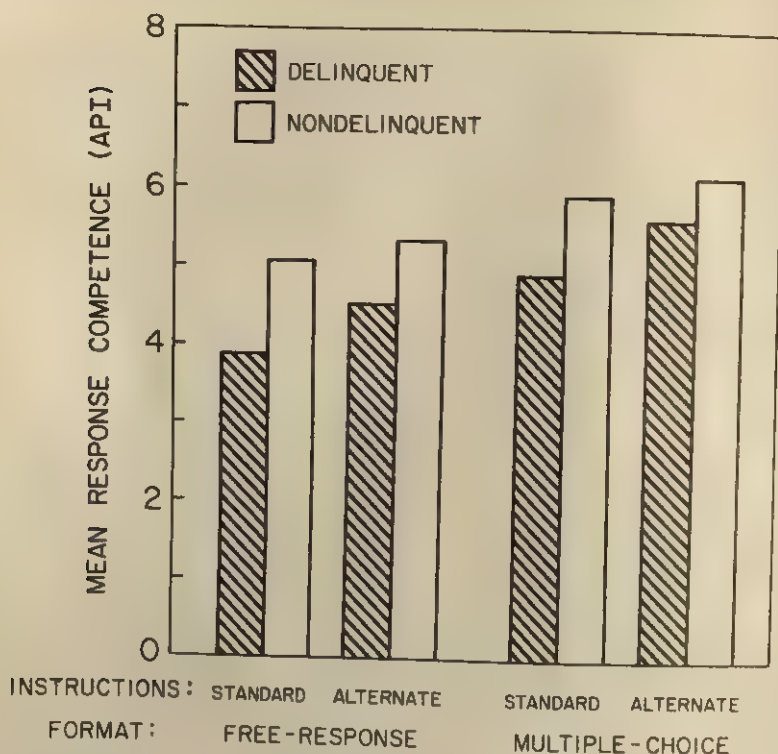


Figure 1. Mean response competence of delinquents and nondelinquents under two sets of instructions (standard vs. alternate) and two test formats (free response vs. multiple choice).

performed significantly better on the multiple-choice version than on the free-response version, $t(39) = 8.86$, $p < .001$. The same pattern was also found with nondelinquents, $t(39) = 3.55$, $p < .001$.

Significant differences were found between delinquents and good citizens on 42 out of 44 items in Study 1. In the present study, however, an item-by-item comparison between mean scores obtained under standard instructions in the free-response format (i.e., the instructions and format used in Study 1) yielded significant group differences ($p < .05$) on only 13 items. This result must be qualified by a number of considerations: The present study was designed with different objectives in mind; the number of subjects contributing to the group mean on each item was only 10, 5 of whom were administered the free-response test following the multiple-choice test; and there was a tendency for subjects to perform better on free-response items after being exposed to the multiple-choice items and answers. Furthermore, re-

sponses in this study were rated from typed transcripts rather than from audiotapes, and this may have eliminated some of the distinguishing response behaviors. It was deemed inappropriate to attempt a more detailed item analysis with the present data, since there were only 5 subjects per cell.

Discussion

The first phase of this research was concerned with the development of a behavioral role-playing inventory of social skills in adolescent boys—the API—and a criterion-referenced raters' manual to accompany the inventory. The second phase was concerned with evaluating the API. Three studies were conducted, along with several psychometric analyses of the inventory itself.

Study 1 found that the API could differentiate between delinquents and nondelinquents, and that nondelinquents who differed in their social competence, according to school guidance counselors' nominations, also differed in

their overall API performance. The API was shown to be reliable; to lack any strong, coherent, or interpretable item structure; and to be an unusually strong predictor of subjects' group membership. Study 2 demonstrated that the API was not merely sensitive to gross differences between delinquent and nondelinquent groups. Institutionalized delinquents who were known to differ in their disruptiveness within the institution also were significantly different in their API performance. This finding suggests that subcultural differences alone probably do not account for differences in API performance.

The third study replicated the previous findings. The API again differentiated between delinquent and nondelinquent subjects. Such differences were found even though several methodological improvements were introduced—namely, in subject-matching and in response-rating procedures. Moreover, the study found that instructions affected the API performances of both delinquents and nondelinquents: When told to respond with "the best" solution, all subjects did better than when they were told to say what they would actually "do." The study also showed that all subjects did better when given the API items in a multiple-choice format than when given the items in a free-response format.

Taken as a whole, the present research attests to the utility of a competence model of assessment. Specifically, the research suggests that the API is a valid measure of social competence in adolescent boys. Generally, the research provides support for the hypothesized relationship between social skill deficits and interpersonal/legal difficulties.

The findings of the present research are consistent with the results of previous studies in the area of delinquency. Numerous investigators have repeatedly demonstrated that delinquents, as a group, do not perform as satisfactorily as nondelinquents on a variety of measures. The critical difference between this research and previous work is not in its demonstration of performance deficiencies among delinquents; rather, the difference is in how it proposes that such deficiencies should be identified, assessed, and interpreted. Previous investigators typically have sought

to isolate either a core cause (e.g., genetic flaw; weak superego) or a unitary personality trait (e.g., aggressiveness) to explain delinquency. The present research, in contrast, was based on the notion that a wide and varied array of skill deficits may be related to delinquent behavior. Both delinquents and nondelinquents are likely to perform competently in some tasks and incompetently in others, and the pattern of particular deficiencies will vary considerably among individuals and within groups. No single deficit or pattern of deficits is likely to explain delinquency. Rather, the probability that an individual will be classified as a delinquent increases as a function of at least three factors: First, it increases to the extent that the individual lacks the requisite skills to deal effectively with the everyday problem situations confronting him; second, it increases as a function of the frequency with which he encounters such problem situations; and third, it increases as a function of the degree to which his incompetent solutions to such problem situations take the form of illegal behaviors.

Although the present research is a first step toward identifying some of the relevant problem situations and common performance deficits characteristic of delinquent boys, a number of important questions remain to be answered by future research.

First, it will be important to gather additional evidence on the external validity of the API; that is, do subjects' role-played responses to the API correlate with their actual responses to the same situations in real life?

Second, far more attention must be devoted to the task of defining situation-specific response competence before we can feel confident about using the API situations and scoring criteria to develop a valid skill-training program for adolescent boys. The criteria for evaluating competence in the present research were based on the judgments of so-called experts. Even though it is encouraging that there was consensus among the judges and that their criteria successfully discriminated among subject groups, it is possible that such criteria simply are not valid when applied to the life contexts of delinquents. A

response that is an effective solution to an interpersonal problem in one milieu may not be very effective in another. Rather than relying on the subjective opinions of judges to define competence, future research should take a more empirical approach; competence in a specific situation should be defined in terms of an objective assessment of the actual consequences of the various response alternatives.

The competence model and the methodology of the present study might profitably be applied to the assessment of other pathological behaviors (e.g., depressive behaviors, alcohol abuse). Gradually, a classification could be developed of particular problem situations and skill deficits that are associated with a variety of clinical populations. Such a classification could be individualized to take into account the specific strengths and weaknesses in the skills repertoire of any new client. Since treatment programs are only as good as the underlying assessment, those of us involved in social skills training might do well to call a moratorium on the further elaboration of treatment programs until we have more systematically assessed both the nature of the clinical problem and the type and extent of the skill deficits of the individuals with whom we are working.

Reference Note

1. Hollingshead, A. B. *Two factor index of social position*. Unpublished manuscript, Yale University, 1957.

References

- Goldfried, M. R., & D'Zurilla, T. A. Behavioral-analytic model for assessing competence. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 1). New York: Academic Press, 1969.
- Goldsmith, J. B., & McFall, R. M. Development and evaluation of an interpersonal skill-training program for psychiatric inpatients. *Journal of Abnormal Psychology*, 1975, 84, 51-58.
- Lachenbruch, P. A., & Mickey, M. R. Estimation of error rates in discriminant analysis. *Technometrics*, 1968, 10, 1-11.
- Lance, G. N., & Williams, W. T. A general theory of classification sorting strategies. I. Hierarchical systems. *Computer Journal*, 1967, 9, 373-380.
- Lord, F. M. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 1967, 68, 304-305.
- McFall, R. M. *Behavioral training: A skill-acquisition approach to clinical problems*. Morristown, N.J.: General Learning Press, 1976.
- McFall, R. M., & Twentyman, C. T. Four experiments on the relative contributions of rehearsal, modeling, and coaching to assertion training. *Journal of Abnormal Psychology*, 1973, 81, 199-218.
- Meehl, P. E., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, and cutting scores. *Psychological Bulletin*, 1955, 52, 194-216.
- Sarason, I. G., & Ganzer, V. J. *Modeling: An approach to the rehabilitation of juvenile offenders* (Final report to the Social Rehabilitation Service, Grant 15-P-55303). Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1971.
- Sneath, P. H. A., & Sokal, R. R. *Numerical taxonomy*. San Francisco: Freeman, 1973.
- Sobell, M. B., & Sobell, L. C. Individualized behavior therapy for alcoholics. *Behavior Therapy*, 1973, 4, 49-72.
- Twentyman, C. T., & McFall, R. M. Behavioral training of social skills in shy males. *Journal of Consulting and Clinical Psychology*, 1975, 43, 384-395.

Received December 29, 1977 ■

Psychosocial Correlates of Empirical Types of Multiple Drug Abusers

G. Nicholas Braucht
University of Denver

Michael W. Kirby
University of Colorado Medical School

G. James Berry

Systems Evaluation and Research Corporation, Denver, Colorado

The present study had three primary focuses: (a) to identify major empirical clusters of multiple drugs; (b) based on these empirical drug clusters, to develop an empirical typology of multiple drug abusers; and (c) to characterize each derived drug cluster and type of multiple drug user in terms of a coherent set of theoretically based psychosocial variables. To accomplish these objectives, drug use and psychosocial data were collected from 440 clients in four drug and alcohol treatment programs. A cluster analysis was performed on chronicity/frequency indices that had been calculated for each of 15 drug classes. Four multiple drug clusters were identified by this analysis: (a) cocaine/other opiates and synthetics/methaqualone/illegal methadone; (b) inhalants/codeine/nonnarcotic analgesics; (c) marijuana/amphetamines/hallucinogens; and (d) minor tranquilizers/barbiturates. Two substances, heroin and alcohol, did not cluster with any other substances but were frequently used by this sample, and consequently these two substances were retained in further analyses, yielding six basic drug clusters. Next, a typology of drug abusers, rather than drug clusters, was developed empirically by means of proximity cluster analysis. Eight quantitatively and qualitatively distinct types of multiple drug abusers were identified solely by analysis of their standing on the use of the six basic clusters of drugs. Finally, the set of psychosocial measures was found to be differentially related to use of the six types of drugs and to the eight types of drug abusers. These differential findings were discussed in terms of the adequacy of the theory underlying the measures and in terms of alternative analytic strategies.

Evidence has accumulated in recent years documenting a substantial increase in the polydrug and multiple drug abuse patterns encountered in a variety of drug treatment settings (Benvenuto & Bourne, 1975; Carlin

& Post, 1971; Duncan, 1975; Fisher & Brickman, 1974; Smith & Wesson, 1973), involving a concomitant proliferation of distinctive multiple drug use patterns, that is, regular use of two or more kinds of drugs. Consequently, there is a pressing need for systematic research into identification of the major types of multiple drug abusers requiring some kind of drug treatment. To develop such a typology, an initial task is to identify major patterns of multiple drug abuse and to describe the psychosocial factors that are associated with each major pattern. The present study has three primary aims: (a) to identify major empirical clusters of multiple drugs; (b) based on these empirical clusters, to develop an empirical typology of

This study was supported in part by funds provided by the Special Action Office for Drug Abuse Prevention and the National Institute of Drug Abuse through the Colorado Department of Health and the Denver Department of Health and Hospitals (Grant 1-H81-DA-01687).

The authors wish to express their gratitude to both agencies and the staffs of the treatment centers involved in the study.

Requests for reprints should be sent to G. Nicholas Braucht, Department of Psychology, University of Denver, Denver, Colorado 80208.

multiple drug users; and (c) to describe each cluster of drugs and type of user in terms of the same set of theoretically derived psychosocial variables.

From a review of the existing research literature in this area, it is our contention that the simultaneous investigation of these three aims would represent a distinct advance over present research strategies. This is so because nearly all of the present literature germane to the above aims reflects one of four approaches:

1. Some investigators have attempted to identify distinct patterns of multiple drug abuse but have provided no information about the distinguishing characteristics of persons using different patterns of drugs (e.g., Simpson & Sells, 1974; Smart & Whitehead, 1972).

2. Others have provided systematic descriptions in terms of theoretically articulated sets of psychosocial variables and their relationships to various forms of social deviance, but these investigators have not focused on patterns of multiple drug use nor have they studied significant numbers of drug abusers with problems of the magnitude typically encountered in drug treatment populations (e.g., Braucht, 1974; R. Jessor, 1976; R. Jessor, Jessor, & Finney, 1973).

3. A third approach is represented by studies concerned with only one pattern of multiple drug abuse, often providing anecdotal or demographic descriptors not based on any identifiable theoretical perspective (e.g., Campbell & Freeland, 1974; Chambers, 1969; Devenyi & Wilson, 1971; Hamburger, 1964; Kirby & Berry, 1975; Ludwig & Levine, 1965).

4. A fourth approach has attempted to identify multiple drug use patterns among general high school or college populations (e.g., Blum, 1969; Brehm & Back, 1968; Groves, 1974; Johnston, 1973) or has attempted to describe initial stages in the progression of drug use among general adolescent populations (Kandel, 1975; Kandel & Faust, 1975; Single, Kandel, & Faust, 1974)—populations without significant numbers of persons having the extensive involvement, experience, and advanced problems with drugs

that constitute the norm among drug treatment clientele.

Against this background of extant research strategies and knowledge, it would be advantageous to develop research that would integrate the best features of these lines of inquiry. Thus, the focuses of the present article are on the identification of distinct types of multiple drug use patterns and multiple drug users in a large sample of persons having major drug problems and on an analysis of the psychosocial characteristics of those types of multiple drug users.

Having stated the objectives and *raison d'être* for the present study, the literature would suggest that an optimal investigation of these objectives must possess several key methodological features. To achieve the first two objectives—identifying major empirical clusters of drugs and types of multiple drug abusers—three requirements must be met: (a) the reduction of a large potential number of drug use patterns to some manageable number without loss of information, a requirement met by any of several forms of dimensional analysis; (b) obtaining a large, maximally heterogeneous sample of persons who are regular users of more than one kind of drug; and (c) assessment of the frequency and chronicity of each kind of drug used for each person in the sample.

To address the third objective—theoretically based differential descriptions of multiple drug use patterns and types of multiple drug abusers—an approach that appears to be promising in accounting for drug abuse must be identified, and appropriate measures derived from the theory must be selected. In the present study, the theoretical approach is a variant of social learning theory as developed by Jessor and his colleagues (R. Jessor, Graves, Hanson, & Jessor, 1968; R. Jessor & Jessor, 1977). This theory has been fruitful in understanding and predicting adolescent deviant behavior, including alcohol and marijuana use. Principal measures in this study included a set of sociocultural and personality variables developed by R. Jessor et al. (1968), R. Jessor, Collins, and Jessor (1972), R. Jessor et al. (1973), R. Jessor (1976), R. Jessor and Jessor (1975), and Braucht (1974). It should be emphasized that the

present study is not intended to be a theory-testing endeavor; rather, as an initial exploratory inquiry, the present study uses these theory-based variables in a purely descriptive role.

Method

Structured interviews were conducted, with 440 clients receiving treatment in seven drug programs located in the Denver metropolitan area. The largest percentage were patients in the Polydrug Treatment Center ($n = 164$); 100 were patients in the Denver General Hospital Alcohol Detoxification Ward; 102 were patients in two methadone programs; and the remaining 74 were patients in three other Denver drug programs. The anonymity of respondents was assured. Across this sample of 440 clients, the average age was 30.5 years, 73% were male, and the ethnic groups represented were as follows: 48% Anglo American, 31% Mexican American; 14% Black-American, and 7% American Indian or Oriental. The socioeconomic level of the sample was predominantly lower middle class. It may give some perspective on the nature of this sample to note that 59% reported having been convicted of at least one (nontraffic-related) crime, whereas 44% reported two or more convictions.

Data collected included age, sex, ethnicity, socioeconomic status level, number of arrests, and number of convictions (lifetime basis, whether drug related or not); a two-factor self-concept scale (Miskimins & Braucht, 1971); a scale of stressful life events (Holmes & Rahe, 1967); a measure of assaultiveness (Buss & Durkee, 1957), and a measure of hostility guilt (Mosher, 1966). The interview also included 34 other psychosocial scales; because these measures and the theoretical framework from which they were derived have been described previously (Braucht, 1974; R. Jessor et al., 1972; R. Jessor et al., 1968; R. Jessor & Jessor, 1975; S. L. Jessor & Jessor, 1974), they will not be elaborated on here.

Many investigators, including Sadava (1975), Walizer (1975), and Gorsuch and Butler (1976), have noted that most studies of drug use have limited value because they have relied either on simple measures of use versus nonuse of a particular drug or have measured drug use along a set of absolute frequency of use categories, ignoring both (a) the length of the period of time during which the drug had been used and (b) the total pattern of multiple polydrug use. The present approach attempted to redress both of these limitations by assessing the use of a number of drugs in terms of both the frequency of use and the length of time that these drug(s) had been used.

Fifteen drug classes were studied: heroin, illegal methadone, other opiates and synthetics, alcohol, barbiturates, amphetamines, cocaine, marijuana, hallucinogens, codeine, nonnarcotic analgesics, volatile inhalants, major tranquilizers, minor tranquilizers, and methaqualone. Each client was questioned extensively as to current and past drug use history for each of the 15 drug classes.

During pilot work with clients of three of the drug programs, it was found that they could reliably provide information about which drugs they had been using on a recent, regular basis, and that they could also reliably report how frequently they used these drugs and how long they had been using them. To incorporate as much of this information as possible in a single index for each of the 15 drugs, a chronicity/frequency (C/F) index was calculated for each drug class used, representing the product of (a) the number of months of recent, regular use of that drug and (b) the number of times per week that the person used the drug across the time in a above. Thus, for example, a drug used either for 6 months on a daily basis (6×7) or for 3 months on a twice daily basis (3×14) yields a C/F index of 42; a drug used daily for 1 year yields a C/F index (12×7) of 84; and a drug used three times per day across a period of 1 year (12×21) yields a C/F index of 252. It should be noted that except for alcohol use, the definition of "regular use" was left to the subject. That is, the interviewer did not attempt to impose a predetermined definition for the amount of a drug that constituted a single use of the substance. With regard to alcohol use, the questions asked specifically about drinking to drunkenness, so that drinking to a drunken state constituted a single use episode for this drug.

Results

Cluster Analysis of Chronicity/Frequency Indices

To identify the most general groups of drugs defined by the 15 C/F drug use indices, a full cycle key cluster analysis was performed (Tryon & Bailey, 1970). Four cluster dimensions were identified involving 11 of the original 15 drugs. The remaining four drugs included (a) heroin, (b) alcohol, (c) major tranquilizers, and (d) the over-the-counter drugs.

Inspection of the mean C/F indices across the entire sample for the latter four drugs showed that the major tranquilizers and over-the-counter drugs were little used (12th and 14th ranked of the 15 drugs), whereas heroin (6th ranked) and alcohol (2nd ranked) were used to a considerable degree. For this reason, both heroin and alcohol were included as the 5th and 6th basic drug clusters in addition to the four identified by the empirical key cluster analysis. Even though these two single substances are not clusters of multiple drugs, it was decided to retain them throughout subsequent analyses in order to provide a com-

parative perspective against which the four true clusters could be viewed.

The resulting six drug substance dimensions (patterns of multiple drug use involving 14 of the original 15 drugs) accounted for 99% of the original communality and 92% of the mean square of the original correlation matrix. Thus, these six clusters of drugs can be studied as they relate to each other and to other variables with very little loss of information. In the section below, each cluster is described in terms of (a) the drugs defining the cluster and (b) statistics indicating the strength of association among usage of the drugs defining each cluster.

Drug Cluster 1: (*cocaine, other opiates and synthetics, methaqualone, and illegal methadone*). The use of the four drug classes comprising this basic drug cluster is highly interrelated—The heavy user of one drug class is very likely to be a heavy user of the other drugs and vice versa. This is evidenced by the domain validity for this first drug cluster of .85 and the cluster score's alpha reliability coefficient of .72. Across the entire sample the average intercorrelation of the C/F drug use scores among these four drugs was .40.

Drug Cluster 2: (*inhalants, codeine, and nonnarcotic analgesics*). The use of inhalants, codeine, and the nonnarcotic analgesics is also strongly related in our sample—Persons who use one heavily are likely to be heavy users of the other two. This drug cluster had a domain validity of .82 and an alpha reliability of .68. Across the entire sample, the average intercorrelation of these three C/F drug use scores was .36.

Drug Cluster 3: (*marijuana, amphetamines, and hallucinogens*). As shown by this third cluster's domain validity (.76) and the cluster score's alpha reliability (.57), the degrees of use of marijuana, amphetamines, and hallucinogens are strongly associated with one another—The heavy user of one is likely to be a heavy user of the others. Across the sample as a whole, the average intercorrelation of the C/F drug use scores for these three drugs was .28.

Drug Cluster 4: (*minor tranquilizers and barbiturates*). The use of minor tranquilizers and barbiturates was significantly as-

sociated in our sample ($r = .36, p < .001$). This drug cluster's domain validity was .62, and its alpha reliability was .38. Thus, the heavy user of barbiturates is also likely to be a heavy user of the minor tranquilizers and vice versa.

As stated earlier, neither heroin nor alcohol formed a multiple drug pattern with one another or with any other drugs. Because of the significant levels of use of both heroin and alcohol in this sample, however, these two drugs have been included as the fifth and sixth basic drugs in the remaining analyses.

Correlations Among the Six Clusters of Drugs

The correlations among the cluster scores for the six patterns of multiple drugs are presented in Table 1, which shows that the six clusters form moderately intercorrelated oblique clusters of multiple drugs.

Relationships Among Use of the Six Basic Patterns of Drugs and Other Drug Use Variables

Table 2 presents correlations among the six basic drug cluster scores and scores on four indices of polydrug use: (a) number of drugs (up to 15 possible) ever used; (b) number of drugs used recently on a regular basis; (c) number of drugs used recently on a daily basis; and (d) extent to which subjects reported the deliberate use of drugs in combinations for the express purpose of achieving some desired effect.

It may give some perspective on the relationships in Table 2 to note that across the entire sample, the mean number of drugs ever used was 6.5 drugs (out of a possible 15), that the average number of different drugs currently used on a regular basis was 2.8 drugs, and that an average of 2.0 different drugs were being used on a daily basis.

Relationships Among Use of the Six Types of Drugs and Psychosocial Variables

Table 3 presents the correlations among use of the six clusters of drug use (based on simple sum, nominally weighted cluster scores) and selected psychosocial variables.

Table 1

Intercorrelations, Internal Consistency, and Mean Use Levels of the Six Empirical Clusters of Drugs

Drug cluster	1	2	3	4	5	6
1. Cocaine/other opiates & synthetics/ methaqualone/illegal methadone	—	.06	.24**	.12*	.10*	.05
2. Inhalants/codeine/nonnarcotic analgesics		—	.15**	.24	-.04	.00
3. Marijuana/amphetamines/hallucinogens			—	.11*	.19**	.05
4. Minor tranquilizers/barbiturates				—	-.03	.07
5. Heroin					—	-.06
6. Alcohol (to drunkenness)						—
Domain validity of drug type cluster	.85	.82	.76	.62	—	—
Alpha reliability of drug type cluster score	.72	.68	.57	.38	—	—
<i>M</i> chronicity/frequency of use	68.8	27.6	293.0	94.8	45.6	118.8

* $p < .05$.

** $p < .01$.

Table 3 shows that the six drug cluster scores are differentially associated with 36 psychosocial variables (7 other variables that were available for this analysis do not appear in Table 3 because correlations with them failed to reach the $p < .01$ level).

First, Table 3 shows that the level of use of the first drug cluster (cocaine/other opiates and synthetics/methaqualone/illegal methadone) is significantly related to only two psychosocial variables. A high level of use of drugs in this first cluster is associated with (a) the presence of a peer subculture that is supportive of such drug use and (b) a high level of life stresses (the only group of multiple drugs to be significantly related to life stress level).

Second, Table 3 shows that the use of

the inhalants/codeine/nonnarcotic analgesics cluster is significantly related to six psychosocial variables. The greater the use of drugs belonging to this second cluster, (a) the less the religiosity, (b) the greater the alienation, (c) the more external the locus of control orientation, (d) the more important the personal effects (escape from problems) functions of such use, (e) the more strained the communication regarding drugs with one's family, and (f) the greater the discrepancy between the kind of person one feels one is versus the kind of person one would ideally like to be. A group picture suggested by this array of relationships is that inhalants/codeine/nonnarcotic analgesics use is associated with a psychosocial syndrome of alienation, lack of values and norms, lack of meaning-

Table 2

Significant Correlations Among the Six Empirical Clusters of Drugs and Four Indices of Polydrug Use

Polydrug index	Cluster					
	1	2	3	4	5	6
No. drugs ever used		.16	.40		.16	-.20
No. drugs used recently on a regular basis	.37	.27	.66	.23	.15	
No. drugs used recently on a daily basis	.96		.13			
Extent of the use of drugs in combination to achieve an effect	-.14	.17	.60	.30		

Note. All table entries are correlations significant at the $p < .01$ level; correlations that failed to attain significance at the $p < .01$ level have been omitted.

ful channels of communication with one's family, feelings of helplessness and inferiority, and a strong desire to escape from the pressure of personal problems.

Inspection of Table 3 shows that the use of the third class of drugs (marijuana/amphetamines/hallucinogens) is related to 25 psychosocial characteristics (presented in the third column of Table 3). The general psychosocial condition suggested by this array of significant relationships is one of marijuana/amphetamines/hallucinogens use on the part of youth embedded in family and peer groups

who model and support drug use. The use of drugs in this cluster appears to be firmly entrenched in a drug subculture with all of its characteristics as popularly conceived. Use of this cluster is not only associated with the desire to escape from personal problems but is also associated with the function of enhancing social occasions, the only class of drugs that is perceived to serve this social enhancement function.

In contrast, Table 3 shows that use of the minor tranquilizers/barbiturates cluster is associated with a pattern of psychosocial char-

Table 3

Significant Correlations Among the Six Drug Clusters and Psychosocial Variables

Psychosocial variable	Cluster					
	1	2	3	4	5	6
Age in years			-.28	.14		.25
Need value for conformance goals				-.19		
Intolerance of deviance			-.17			
Religiosity		-.13	-.25	-.17		
Need value for conformance success			-.16	-.19		
Alienation		.18	.25	.17		
Social agents' agreement			-.17			
Life chances disjunctions				.15		
Peers' advice salience			-.17	-.12		
Expectation of conformance success				-.23		
Family advice salience			-.21			
Peers' value for conformance goals			-.15	-.15		
Internal locus of control		.16	.14	.24		
Peers' value for conformance goals			-.17			
Opportunity to procure drugs			.19	.20		
Peer support for drug use	.16		.42			
Positive social function of drugs			.23			
Personal effects function of drugs		.17	.24	.24		
Conforming social functions of drugs						.13
Family value for conformance goals					.14	
Exposure to parental medical drug use			.28			
Family support for drug use			.32			
Family proscription of alcohol use			.14			
Internal negative functions of drug use			.12			
External negative functions of drug use			.15			
Hostility guilt score			-.15			
Peers' sanction likelihood			-.20			
Ease of communication re drugs with family						
Ease of communication re drugs with peers		-.14				
Assaultiveness			.14			
Total life stress score			.13	-.24		
No. arrests	.13					
No. convictions				-.16	.17	.18
Socioeconomic status						.29
Self/ideal-self discrepancy						-.20
Self-others, over valuing others		.14	.19	.19		.19

Note. All table entries are correlations significant at the $p < .01$ level; correlations that failed to attain significance at the $p < .01$ level have been omitted.

Table 4
Drug Use Cluster Score Profiles of the Eight Types of Drug Users

Type	n	%	Profile						Overall homogeneity
			1	2	3	4	5	6	
1. Infrequent experimenters	230	55.0	48.0	48.5	45.8	47.1	48.1	45.3	.99
2. Alcoholics	67	16.0	47.9	48.7	46.9	47.2	47.5	71.0	.96
3. Minor tranquilizer/barbiturate users	17	4.1	49.9	49.0	47.3	80.0	47.5	46.7	.90
4. Heavy heroin and some cocaine/other opiates (no illegal methadone users)	16	3.8	50.9	48.4	50.3	47.5	101.2	45.2	.91
5. Hallucinogen and minor tranquilizer barbiturate users	11	2.6	48.0	48.8	60.4	60.9	47.8	46.6	.86
6. Heavy minor tranquilizer/barbiturate and some codeine/inhalant/analgesics users	8	1.9	50.0	54.0	45.2	58.6	48.4	46.6	.93
7. Codeine (heavy), inhalant/analgesics and some minor tranquilizer/barbiturate users	7	1.7	49.6	77.0	45.4	52.5	47.5	44.7	.89
8. Hallucinogen users	62	14.9	48.5	48.6	61.2	47.3	48.8	45.6	.92
Eta			.98	.98	.91	.98	.97	.97	

acteristics suggestive of older users who are intolerant regarding deviance from conformist ethos, who have pervasive feelings of helplessness in controlling their own destiny, and who lack confidence in their own vocational career success.

The use of heroin is related to only two psychosocial variables: Heavy heroin use is associated with having been arrested several times and with a family ethos that places a high value on the importance of conformist goals.

Finally, Table 3 shows that the use of the last basic type of drug (alcohol) is related to older user age, importance attached to using alcohol as a means of conforming to social expectations, elevated levels of criminal arrests and convictions, low socioeconomic status, and a classic sign of depression—the feeling that others think more of one than one thinks of oneself.

Identifying Distinct Types of Drug Abusers

Thus far, six drug clusters (i.e., six types of *drugs*) have been identified, and the patterns of relationships of each cluster of drugs with a set of psychosocial characteristics have been described. At this point, typology of

drug *abusers*, rather than a typology of drugs, was developed empirically by means of proximity cluster analysis. Eight quantitatively and qualitatively distinct types of multiple drug abusers were identified using the standard procedures of Tryon and Bailey's (1970, p. 147) method of iterative condensation on centroids. Of the 440 abuser profiles, 418 were classified into one (and only one) of the eight drug abuser types. The remaining 22 abusers' profiles were too discrepant or unique to be included in any of the eight types.

Thus, independent of their standing on psychosocial variables, eight distinct types of drug users were identified solely by analysis of their standing on the use of the six basic clusters of drugs. Table 4 provides a rough descriptive label for each type of drug user and presents the distinctive profile of the use of the six basic clusters of drugs for each type of user.

Eta statistics in the bottom row of Table 4 indicate the degree to which each of the six drug clusters account for type membership. Conversely, the eight homogeneity statistics indicate the average degree to which membership in each of the eight drug user types specifies scores on the six drug use cluster

score dimensions. All of these statistics are at highly acceptable levels, indicating that the typology obtained possesses a high degree of

statistical integrity and "tightness" (Tryon & Bailey, 1970, p. 161).

Table 5 presents a more complete descrip-

Table 5

Drug User Types: Significant One-Way Analyses of Variance of Psychosocial Dependent Variables

Variable	Type							
	1	2	3	4	5	6	7	8
<i>n</i>	230	67	17	16	11	8	7	62
total sample percentage	55	16.0	4.1	3.8	2.6	1.9	1.7	14.9
Chronicity/frequency indices of drug use								
Heroin			—	+				—
Methadone		—	—	+	—	—	—	
Opiates				+				
Alcohol	—	+						
Barbiturates	—	—	+					—
Amphetamines	—	—	—	—	+	—	—	—
Cocaine	—	—	—	+	—	—	—	—
Marijuana	—	—	—		+	—	—	+
Hallucinogens			—	+		—		+
Codeine	—	—	—	—			+	—
Analgesics	—	—	—	—	—		+	—
Inhalants	—	—	—	—	—	—	+	—
Tranquilizers								
Major	—	—	—	—	+	—	—	—
Minor	—	—	+	—				—
Methaqualone	—	—	—	—	—	+	+	—
No. drugs ever used		—			+			
No. drugs used recently and regularly	—				+			
No. drugs used daily	—				+			
Sum chronicity/frequency	—				+			
Use of drugs in combination	—			+	+	—		
No. arrests	—	—			+	+		
No. convictions	—	+	—	+	—	—		—
Socioeconomic status	—	+	—	+	—	—		—
Self-others, overvaluing others	—	—					+	
Age	—	+		—				
Sex		— ^a	— ^a		— ^b			— ^b
Male								
Female							All	
Ethnicity		65%						
			88%	— ^d	— ^e	— ^e	— ^e	
			white ^a					
Alienation	—	—	—		+		—	—
Internal locus of control	+	—	—					
Opportunity to procure drugs	—	—	+	—		—	—	
Positive social functions				—	+			+
Conforming social functions		+		—	+	—		
Personal effects functions	—		—	—	+			
Internal negative functions of drug use		—				+		
Assaultiveness			—	+	+	—		+
Peers' value for conformance goals	+	+	+	—			+	
Peers' value for deviant goals	—	—	—			—	—	+

^a Oldest.

^b Youngest.

^c No blacks.

^d Mexican Americans, few whites.

^e No Mexican Americans.

tion of each type, both in terms of drug use patterns and psychosocial characteristics. With drug user type as the independent variable, a series of one-way analyses of variance was performed for each psychosocial variable. Each row of Table 5 represents a significant F ratio of chi-square ($p < .01$) that indicates a significant difference among the eight drug user types on that psychosocial characteristic. The plus and minus entries in each row of Table 5 describe the results of a posteriori multiple-range tests (by the method of least squared differences) that were performed following the significant F test. On any given psychosocial variable, drug user types with a plus sign are relatively high on that characteristic and are significantly different ($p < .01$) from user types with no table entry (who are "average" on that characteristic), who are, in turn, significantly different ($p < .01$) on that psychosocial variable from user types with a minus sign.

Table 5 has been designed to communicate a wealth of significant findings about the characteristics of each type of drug user as directly and completely as possible. The characteristics of each type of drug user suggested by the pattern of findings in Table 5 are briefly summarized below.

Drug User Type 1: Infrequent experimenters. Comprising the bulk of our sample, this type of user has experimented with over 6 different drug classes (of 15 possible) during his/her lifetime and has recently used 1.7 drugs on a regular basis, but is unlikely to have used *any* on a daily basis. This type tends not to use drugs in combinations very often—42% have never consciously used drugs in combinations to achieve an effect.

In contrast to the other types of drug users, these infrequent experimenters feel that they themselves are in control of the direction on their lives. In general, they, their families, and their peer groups are relatively conformist in orientation, and they enjoy a relatively high socioeconomic status. Finally, the infrequent experimenters tend not to use drugs to escape the press of personal problems.

Drug User Type 2: Alcoholics. The alcoholic type of drug user is older, poorer, and

has been convicted of crimes more often (11 times) than any other type of drug user. He or she has had experience with fewer drugs during his/her lifetime than any other type of user (on the average, about three different drugs). The alcoholic type was currently using 1.9 drugs on a regular basis and 1.1 drugs (alcohol) on a daily basis. More than any other type, alcoholics tend *not* to use drugs in combinations, at least not intentionally—65% reported "never" using drugs in combinations to achieve an effect. Alcoholics are relatively conformance oriented in their own values and beliefs, and although they perceive little support for drug use among their peers, they report drinking as one means of conforming to their social group. Compared to any other type of drug user, they recognize fewer drawbacks to their own use of drugs.

Drug User Type 3: Barbiturate and minor tranquilizer users. The user of barbiturates and minor tranquilizers is likely to be an older, white, middle-class housewife who feels powerless to control the direction of her own life. Relatively conformist in orientation, and with the lowest rate of arrests and convictions, the users of barbiturates and minor tranquilizers were the only type of drug users in our sample who had never had *any* experience with heroin or methadone. In regard to the conscious use of drugs in combinations to achieve an effect, 31% of these barbiturate/minor tranquilizer users reported "always" using such combinations. Also, they reported their access to drugs to be easier than any other type of drug user.

Drug User Type 4: Narcotics users. These users of heroin, illegal methadone, other opiates, cocaine, and (in conjunction with these narcotics) hallucinogenic drugs, tend to hold beliefs contrary to the conformist ethos—They are assaultive and tolerant of deviance. The average narcotics user in our sample had been arrested an average of 28 times but had only been convicted three times. They feel that their peers devalue conformance success more than does any other type of drug user. Their use of these "hard" drugs is heavy, and 31% of this type of user report always using drugs in combinations to achieve an effect. In our sample, 57% of

the members of this type were of Mexican-American descent, the only type composed primarily of this ethnic group.

Drug User Type 5: Amphetamines, marijuana, and major tranquilizer users. Users of amphetamines, marijuana, and the major tranquilizers have used more different drugs during their lives (11 drugs); they have used the widest variety of drugs recently on a regular basis (7 drugs); they have used the most different drugs on a daily basis (3.8 drugs); and they are more likely than any other type of user to be using drugs in combination to achieve an effect—73% of these users reported always using such combinations; and none reported “never” using combinations.

These heavy multiple drug users are extremely young, alienated, and assaultive. They are more likely than any other type of user to value drugs as a means of escaping from their personal problems, and they also see drugs as social enhancers and as means of conforming to their social group. In our sample, no member of this drug user type was of Mexican-American descent.

Drug User Type 6: Methaqualone, minor tranquilizer, and barbiturate users. These drug users feel themselves to be in a highly stressful environment while feeling relatively helpless to change the course of their lives. Very unlikely to be extraverted or assaultive, they do not view drugs as a means of “going along” with the group. In fact, they feel that their peer group disapproves of drug use. These users see more drawbacks to their own use of drugs—in terms of loss of self-control, self-respect, and loss of friends—than does any other type of drug user.

Drug User Type 7: Codeine, inhalant, analgesic, and methaqualone users. This type of user tends to use alcohol (to drunkenness) very infrequently, if at all. They are relatively high in socioeconomic status. In our sample, there were no black members of this drug user type, and all members of this type were male. In general, members of this type hold relatively conformist values and attitudes. Their drug use occurs despite their report that (a) they find drugs hard to get and (b) their peers do not support drug use.

Drug User Type 8: Hallucinogenic users.

Primarily users of marijuana and the hallucinogenic drugs (they use more than any other type of drug user), they also are using some amphetamines. These are young, assaultive users, and in contrast to members of Drug User Type 5 (who use many of the same kinds of drugs), they do not feel alienated. More than any other type of user, they reported that their peers are supportive of drug use, and they see the primary function of drug use to be as a means of enhancing social occasions.

Discussion

To return to the three primary aims of this study, six basic classes of drugs were identified empirically. In the present sample, four multiple drug clusters were found, and neither the use of heroin nor alcohol was strongly associated with the use of any other drug.

In general, the four multiple drug clusters found here do not correspond to patterns previously reported in Simpson and Sells's (1974) investigation of the problem of identifying multiple drug abuse patterns among a drug treatment population. In their study of 11,380 patients who were included in the initial 2 years of the National Institute of Mental Health—Texas Christian University Drug Abuse Reporting Program (DARP), 28 distinctive patterns were found. The most frequent pattern, representing over 28% of the patients, was the daily or weekly use of heroin alone. Moreover, daily or weekly use of heroin with cocaine, marijuana, and with both also had prevalent patterns. These four patterns—each involving heroin—accounted for just over 52% of their patients.

There is, however, a major difference between the Simpson and Sells (1974) study and the present one—the differing requirements for admission to the participant treatment programs. The treatment agencies within DARP were primarily treating opiate addicts, and during the first 2 years, users of other drug classes usually were not admitted unless some level of opiate use was also indicated. Thus, given these selection criteria, the finding of a high prevalence of heroin use, singly and in combination, would be

expected. In contrast, the present study contained subjects from a variety of programs, and although 23% were drawn from an alcohol treatment program and 23% from a heroin treatment program, the majority (54%) were from programs not having these selective admission requirements.

In exploring our second objective—the development of an empirical typology of multiple drug abusers—eight distinct types were identified. These types were an intriguing finding, as some of the types, such as Type 1 (marijuana/hallucinogen users), conform to previously reported and commonly described types, whereas other types would not be expected on an *a priori* basis. For example, Type 5 (amphetamine/marijuana/major tranquilizer users) and Type 7 (codeine/nonnarcotic analgesic/inhalant/methaqualone users) are types that are not widely described in the literature.

As with any study of this sort, it should be noted that the typological findings obtained here are functions not only of “real” drug-using phenomena but also of the “method” used to observe those phenomena. Strictly speaking, the present typology must be regarded as a function of the particular drug use measures used, the composition of the sample studied here, and the specific statistical typing procedure used. In this regard, however, the finding of significant degrees of communality between the key C/F measures and other indices of polydrug use (see Table 2), the achievement of a large sample from a variety of drug program sources, and the selection of a well-known standardized set of cluster analysis procedures all suggest that the obtained typology is relatively robust with regard to these method factors.

The third objective was to determine the differential association of a coherent set of psychosocial variables with each of the six drug clusters and the eight drug user types. The particular set of psychosocial measures used in this study is based on social learning theory as articulated by the Jessors and their colleagues, and these measures have been shown previously to be powerful predictors of adolescent deviant behavior. Because the users of the drugs comprising the

third drug cluster (marijuana/amphetamines/hallucinogens) most closely approximate the subject population on which this theoretical approach has been developed, the strong associations of these psychosocial variables with this drug cluster were not unexpected. Nevertheless, if the explanatory scope of the underlying theory were adequate, one would have expected stronger associations with the remaining multiple drug clusters. The general absence of significant relationships between these psychosocial measures and the other multiple drug clusters (particular with Clusters 1 and 5), as well as the striking differences in strength of association, raises questions about the utility of this theoretical approach in explaining a wide variety of multiple drug use patterns. It may be that alternative theories and their associated constructs will be required to provide sets of descriptive variables that relate to differing drug use patterns and users, as has been implied by several recent articles (see Bentler & Eichberg, 1975; Braucht, Brakarsh, Follingstad, & Berry, 1973; Lettieri, 1975; McGlothlin, 1975). On the other hand, it may be that merely introducing key moderator variables or the use of alternative analytic strategies such as those suggested by Gorsuch and Butler (1976) and Dunnette (1975) will reveal additional explanatory power now latent in the present social learning theory.

Although the present study has succeeded in examining specific subgroups of people in relation to specific types of drugs (advocated by Kessler, Paton, & Kandel, 1976), it should be stressed that this study represents an exploratory effort. Future research is needed to build on this step, replicating and extending the present investigation with larger numbers of subjects from differing locales and differing subcultural groups. To provide perspective, a sample of “normals” should be interviewed using the same set of measures employed with drug-abusing groups. There is also a need for further sampling of theories and theoretical constructs in order to provide an arena for the comparative assessment of competing theories in explaining various patterns of multiple drug abuse.

If this line of research were to be pursued successfully, we believe that the resulting

typology would not only contribute to a basic understanding of these problems, but it would also be of great value to those charged with the prevention and treatment of these problems. Given a comprehensive, general typology, *differential* public health programming could be developed for each type of user; the planning, locating, and staffing of these programs could be expected to benefit from the knowledge base provided by a comprehensive typology. Finally, given such a typology, it should be clear that differential evaluations of prevention and treatment programs could be done in a way that is not now possible—Type of program, type of drug user, and their interaction could be examined. Thus, both in terms of pure understanding and practical utility, additional research along the lines of the present study is needed.

References

- Bentler, P. M., & Eichberg, R. H. A social psychological approach to substance abuse construct validity: Prediction of adolescent drug use from independent data sources. In D. J. Lettieri (Ed.), *Predicting adolescent drug abuse: A review of issues, methods and correlates* (National Institute on Drug Abuse, Research Series No. 11, DHEW Publication No. ADM 76-249). Washington, D.C.: U.S. Government Printing Office, 1975.
- Benvenuto, J., & Bourne, P. G. The federal polydrug abuse project: Initial report. *Journal of Psychotic Drugs*, 1975, 7, 115-120.
- Blum, R. H. *Students and drugs; Drugs II: College and high school observations*. San Francisco: Jossey-Bass, 1969.
- Braucht, G. N. A psychosocial typology of adolescent alcohol and drug users. In M. H. Chafetz (Ed.), *Proceedings of the Third Annual Alcoholism Conference of the National Institute on Alcohol Abuse and Alcoholism* (DHEW Publication No. ADM 75-137). Washington, D.C.: U.S. Government Printing Office, 1974.
- Braucht, G. N., Brakarsh, W. D., Follingstad, D., & Berry, K. L. Deviant drug use in adolescence: A review of psychosocial correlates. *Psychological Bulletin*, 1973, 79, 92-196.
- Brehm, M. L., & Back, K. W. Self-image and attitudes toward drugs. *Journal of Personality*, 1968, 36, 299-314.
- Buss, A. H., & Durkee, A. An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology*, 1957, 21, 343-349.
- Campbell, R. S., & Freeland, J. B. The hippie turns junkie: The emergence of a type. *International Journal of the Addictions*, 1974, 9, 719-730.
- Carlin, A. S., & Post, R. D. Patterns of drug use among marihuana smokers. *Journal of the American Medical Association*, 1971, 218, 867-868.
- Chambers, C. D. Barbiturate-sedative abuse: A study of prevalence among narcotic abusers. *International Journal of the Addictions*, 1969, 4, 45-57.
- Devenyi, P., & Wilson, M. Abuse of barbiturates in an alcoholic population. *Canadian Medical Association Journal*, 1971, 104, 219-221.
- Duncan, D. F. The acquisition, maintenance and treatment of polydrug dependence: A public health model. *Journal of Psychedelic Drugs*, 1975, 7, 209-213.
- Dunnette, M. D. Individualized prediction as a strategy for discovering demographic and interpersonal/psychological correlates of drug resistance and abuse. In D. J. Lettieri (Ed.), *Predicting adolescent drug abuse: A review of issues, methods and correlates* (National Institute on Drug Abuse, Research Series No. 11, DHEW Publication No. ADM-76-240). Washington, D.C.: U.S. Government Printing Office, 1975.
- Fisher, G., & Brickman, H. R. Multiple drug use of marihuana users. *Diseases of the Nervous System*, 1974, 34, 40-43.
- Gorsuch, R. L., & Butler, M. C. Initial drug abuse: A review of predisposing social psychological factors. *Psychological Bulletin*, 1976, 83, 120-137.
- Groves, W. E. Patterns of college student drug use and life-styles. In E. Josephson & E. Carroll (Eds.), *Drug use: Epidemiological and social approaches*. New York: Wiley, 1974.
- Hamburger, E. Barbiturate use in narcotic addicts. *Journal of the American Medical Association*, 1964, 189, 366-368.
- Holmes, T. H., & Rahe, R. H. The Social Readjustment Rating Scale. *Journal of Psychosomatic Research*, 1967, 11, 213-218.
- Jessor, R. Predicting time of onset of marijuana use: A developmental study of high school youth. *Journal of Consulting and Clinical Psychology*, 1976, 44, 125-134.
- Jessor, R., Collins, M. I., & Jessor, S. L. On becoming a drinker: Social psychological aspects of an adolescent transition. In F. A. Seixas (Ed.), *Nature and nurture in alcoholism*. New York: Annals of the New York Academy of Sciences, 1972.
- Jessor, R., Graves, T. D., Hanson, R. C., & Jessor, S. L. *Society, personality and deviant behavior: A study of a tri-ethnic community*. New York: Holt, Rinehart & Winston, 1968.
- Jessor, R., & Jessor, S. L. Adolescent development and the onset of drinking: A longitudinal study. *Journal of Studies on Alcohol*, 1975, 36, 27-51.
- Jessor, R., & Jessor, S. L. *Problem behavior and psychosocial development*. New York: Academic Press, 1977.
- Jessor, R., Jessor, S. L., & Finney, J. A social psychology of marijuana use: Longitudinal studies of high school and college youth. *Journal of Personality and Social Psychology*, 1973, 26, 1-15.
- Jessor, S. L., & Jessor, R. Maternal ideology and adolescent problem behavior. *Developmental Psychology*, 1974, 10, 246-254.

- Johnston, L. *Drugs and American youth*. Ann Arbor, Mich.: Institute for Social Research, 1973.
- Kandel, D. B. Stages in adolescent involvement in drug use. *Science*, 1975, 190, 912-914.
- Kandel, D. B., & Faust, R. Sequence and stages in patterns of adolescent drug use. *Archives of General Psychiatry*, 1975, 32, 923-932.
- Kessler, R. C., Paton, S. M., & Kandel, D. B. Reconciling unidimensional and multidimensional models of drug use. *Journal of Studies on Alcohol*, 1976, 37, 632-647.
- Kirby, M. W., & Berry, G. J. Selected descriptive characteristics of polydrug abusers. *Journal of Psychodetic Drugs*, 1975, 7, 161-167.
- Lettieri, D. J. (Ed.). *Predicting adolescent drug abuse: A review of issues, methods and correlates* (National Institute on Drug Abuse, Research Series No. 11, DHEW Publication No. ADM 76-249.) Washington, D.C.: U.S. Government Printing Office, 1975.
- Ludwig, A. M., & Levine, J. Patterns of hallucinogenic drug abuse. *Journal of the American Medical Association*, 1965, 191, 104-108.
- McGlothlin, W. H. Drug use and abuse. *Annual Review of Psychology*, 1975, 26, 45-64.
- Miskimins, R. W., & Braucht, G. N. *Description of the self*. Fort Collins, Colo.: Rocky Mountain Behavioral Science Institute, 1971.
- Mosher, D. L. The development and multitrait-multimethod matrix analysis of three aspects of guilt. *Journal of Consulting Psychology*, 1966, 30, 25-29.
- Sadava, S. W. Research approaches in illicit drug use: A critical review. *Genetic Psychology Monographs*, 1975, 87, 3-59.
- Simpson, D. D., & Sells, S. B. Patterns of multiple drug abuse: 1969-1971. *International Journal of the Addictions*, 1974, 9, 301-314.
- Single, E., Kandel, D. B., & Faust, R. Patterns of multiple drug use in high school. *Journal of Health and Social Behavior*, 1974, 15, 344-357.
- Smart, R. G., & Whitehead, P. C. A typology of high school drug use: Medicinal usage, mood-modification, and tripping. *International Journal of the Addictions*, 1972, 7, 735-738.
- Smith, D. E., & Wesson, D. R. (Eds.). *Uppers and downers*. Englewood Cliffs, N.J.: Prentice-Hall, 1973.
- Tryon, R. C., & Bailey, D. E. *Cluster analysis*. New York: McGraw-Hill, 1970.
- Walizer, D. G. The need for standardized, scientific criteria for describing drug-using behavior. *International Journal of the Addictions*, 1975, 10, 921-936.

Received December 29, 1977 ■

Comparative Evaluation of Therapeutic Components Associated With Behavioral Marital Treatments

Gayla Margolin

Department of Education
University of California, Santa Barbara

Robert L. Weiss

University of Oregon

This study investigated the relative effectiveness of three therapeutic components common to behavioral marital therapies: procedures designed to change behavior, procedures to change attitudes, and nonspecific therapeutic effects. A hierarchical ordering of these components produced three treatment conditions—nonspecific, behavioral, and behavioral-attitudinal. Twenty-seven couples experiencing marital distress were randomly assigned to one of the three treatment conditions and one of five paraprofessional counselors. After four therapy sessions, the groups were compared on measures of self-reported satisfaction, daily reported pleasing (or displeasing) relationship events, and observations of communication skillfulness. All groups showed significant decreases in negative relationship behaviors. The behavioral-attitudinal group, compared to the other groups, showed significantly greater improvement in reported marital satisfaction, pleasing behaviors, and positive communication responses.

Comprehensive reviews on the outcome of marital therapies present evidence supporting the efficacy of behavioral strategies to improve marital functioning but caution that much of the evidence is not based on rigorous empirical investigation (Gurman & Kniskern, *in press*; Jacobson, 1978a). The consistently positive findings on behavioral marital therapy come from series of uncontrolled case studies (e.g., Weiss, Hops, & Patterson, 1973) and, more recently, from studies comparing behavioral marital therapy to no-treatment controls (e.g., Jacobson, 1977). Studies comparing behavioral marital therapy to other theoretical approaches provide

more equivocal results (e.g., Liberman, Levine, Wheeler, Sanders, & Wallace, 1976). In both controlled and comparative studies, the utility of behavioral marital approaches is evidenced more consistently on behavioral outcome measures than on traditional self-report measures (e.g., Harrell & Guernsey, 1976).

Overall, the recent proliferation of outcome studies supports the use of behavioral strategies to increase marital accord, but the studies do not identify which ingredients of the behavioral treatment package are responsible for change. For the most part, investigators have applied multiple intervention procedures without using experimental designs that permit identification of the efficacious treatment components. According to Jacobson and Martin (1976), communication skill training and contingency contracting are the two therapeutic elements most commonly used in behavioral marital therapies. However, recent investigations by Jacobson (*in press*) and by Turkewitz and O'Leary (Note 1) suggest that communication training, by itself, may be sufficient to produce marital improvement. The present study was designed as a components analy-

This manuscript was based on the first author's doctoral dissertation, conducted as partial fulfillment of the requirement for the PhD degree at the University of Oregon. The authors thank Ray Hyman, Miriam Johnson, and Edward Lichtenstein for their recommendations and support.

Portions of this article were presented at the meeting at the Western Psychological Association, Los Angeles, April 1976.

Requests for reprints should be sent to Gayla Margolin, who is now at the Psychological Research and Service Center, 734 West Adams Boulevard, University of Southern California, Los Angeles, California 90007.

sis to examine the relative efficacy of therapeutic ingredients associated with communication training in behavioral marital therapy.

The present study served to isolate two features that are typically included in behavioral marital therapies but are not identified as the components that promote therapeutic change. A nonspecific therapy group was included to isolate and control for nonspecific therapeutic effects that accompany and perhaps enhance behavior change technologies. In addition to controlling for standard nonspecific effects, such as relationship and expectancy factors, this nonspecific group also controlled for the unintentional effects related to behavioral assessment procedures. A recent study by Jacobson (1978b) is the only example in which improvement in behavioral therapy clearly is not a function of nonspecific effects: He found that couples who received a behavioral treatment improved significantly more on three out of four outcome measures than couples in a nonspecific control group.

The present study also isolated cognitive restructuring as a potentially active component of behavioral marital approaches. A behavioral-attitudinal treatment was designed that intentionally included procedures to change attitudes as well as procedures to change behaviors. The purpose of the cognitive restructuring procedures was to help spouses reattribute the source of their marital problems to "bad relationship skills" rather than "bad people" in the relationship. When spouses attribute marital difficulty to the personality of the other spouse, they tend to discount that person's attempts to be reinforcing and maintain their pessimistic impressions regarding the relationship (Margolin, Christensen, & Weiss, 1975). The cognitive restructuring procedures in this study were designed to help spouses to (a) be less blaming; (b) become more accepting and responsive to their partners' efforts to be pleasing; and (c) realize that the goal of relationship improvement was mutual, rather than personal, gain. Rather than being a substitute for behavior change procedures, the cognitive restructuring procedures were intended to make it easier for spouses to

engage in behavior change without losing face.

The three treatment conditions in this study systematically introduced therapeutic components to determine the most efficacious combination of components: (a) nonspecific effects (NS condition); (b) behavioral training plus nonspecific effects (BT condition); and (c) attitudinal restructuring plus behavioral training and nonspecific effects (AB condition). It was predicted that couples in the AB condition would experience greater behavioral improvements in their relationships than those receiving the more standard BT treatment. It was also predicted that couples in treatment groups based on social learning theory (AB and BT) would experience greater behavioral improvements in their relationships than couples in the NS group. Since nonspecific effects have been related to more superficial, as opposed to behavioral, changes, it was hypothesized that couples receiving that treatment would make gains, perhaps even equal to the other groups, on the measures indicating self-reported marital well-being.

Method

Design

This research is best conceptualized as an analogue outcome study, since the treatment offered was an intensified, abbreviated, and standardized version of naturalistic therapy. Treatment parameters not salient to the behavioral or cognitive components were constant across treatment groups, for example, number of sessions and amount of therapist contact. Spouses were seen conjointly at weekly intervals by their own individual therapist over a period of 4 weeks for a total of four 2-hour sessions. The fixed treatment sequence consisted of a (a) 1-week baseline evaluation; (b) 2-week intervention phase; and (c) 1-week posttreatment evaluation. Each couple was randomly assigned to one of the three treatment conditions and to one of five paraprofessional counselors for 4 weeks of couples counseling.

Subjects

Twenty-seven couples experiencing marital distress participated in the study. Although couples were recruited through publicity on radio, television, and in the local newspapers, acceptance into this project was based on evidence of marital distress. Marital distress was initially assessed by the experimenter in a tele-

phone interview and a 1-hour intake session. In addition, couples had to meet two of the following three criteria to be accepted into the project: (a) a mean spouse Locke-Wallace score < 100 ; (b) a total couple Areas of Change score > 15 major problem items; and (c) a therapist rating < 4 based on an 8-point rating scale ranging from high marital distress to high marital adjustment. The mean spouse Locke-Wallace score of participating couples was 71.8 ($SD = 22.5$). Of the 31 couples who volunteered for the project and who completed screening procedures, 2 couples did not meet the criteria for marital distress, 1 couple declined treatment, and 1 refused treatment after the first session. Upon entering counseling, all couples paid a \$50 refundable security deposit that was contingent on keeping scheduled appointments and completing homework assignments. The total deposit was refunded to all except 1 of the 27 participating couples.

Client couples varied substantially in age (range from 20 to 72 years; $Mdn = 31.5$) and income (range from \$3,500 to \$27,000; $Mdn = \$15,000$), and had a relatively high education level. (For 19 couples, one or both spouses had obtained a college degree.) Seven of the couples were childless, but overall the sample averaged 1.4 children per family. For 11 of these couples, one or both partners had been engaged in prior therapy. Analyses of variance were used to examine possible group differences in age, length of marriage, number of children, and income; no significant between-group differences were found.

Therapists

Of the five persons serving as paraprofessional therapists,¹ two had completed college and three were undergraduates. Since none of the therapists had previous experience in counseling couples, they all underwent 40 hours of training prior to seeing their first cases and continued to participate in both group and individual supervision during the time they saw cases. All training and supervision was conducted by the first author.

Measures of Treatment Validity

Two measures were designed to assess the credibility of the major assumptions behind the experimental design that (a) treatment modalities were equally believable and (b) nonspecific couples without specific training would not be knowledgeable about principles of social learning theory applicable to marital problems. A posttest treatment evaluation questionnaire, designed to measure satisfaction with treatment, tested the first assumption. The Consultation Readiness Inventory, consisting of multiple-choice questions following brief vignettes of marital conflicts (similar to the Inventory of Marital Conflicts by Olson & Ryder, 1970), was used to measure spouses' knowledge of and capacities for applying behavioral principles. Spouses independently read the vignettes of hypothetical marital problems and identified the statement that best described factors contributing to each problem and/or strategies for remedying each problem.

Measures of Treatment Outcome

Multidimensional outcome criteria were used to determine how the separate therapeutic components contribute differentially to specific outcomes. Six distinct measures were used: three self-report inventories, two daily reported observational indices from the home, and observational data from videotaped interactions in the laboratory.

Self-report satisfaction inventories. Three questionnaires were used to measure global marital satisfaction of spouses prior to the first session and after the final session. The Locke-Wallace Marital Adjustment Inventory (Locke & Wallace, 1959), which was chosen for its wide usage in traditional marital literature, provides an index of marital compatibility and stability. The Areas of Change Questionnaire (Weiss et al., 1973) provides an index for measuring the amount of change that each spouse desires in the partner's performance of specific relationship behaviors. This 34-item inventory asks each respondent to indicate on a 7-point scale whether (a) she/he desires the spouse to change a particular behavior and (b) it would please the partner if the respondent were to change. Each spouse also completed the Adjective Check List (Gough & Heilbrun, 1965) indicating which adjectives accurately described the partner, thereby capturing the evaluative trait labels that spouses cognitively assign to one another.

Daily reported home observations. Every evening throughout the 28 days of treatment, each spouse recorded the number of pleasing and displeasing relationship events that she/he received from the partner as well as the frequency of pleasant thoughts that she/he had about the partner. Pleasing and displeasing behavioral events were recorded by means of the Spouse Observation Checklist (SOC; Weiss et al., 1973; Weiss & Margolin, 1977), which provides 400 sample relationship behaviors representative of 12 areas of marital functioning (e.g., communication, affection, companionship). The specific behaviors have either a pleasing (e.g., "Spouse gave me a massage") or displeasing (e.g., "Spouse criticized me") effect on the recipient. Each spouse also recorded the occurrence of pleasant thoughts about the partner on a tracking card.

Laboratory observational data. During the first and last sessions, each couple engaged in two 10-minute negotiation sessions during which the spouses attempted to solve a relationship problem without any interruptions or assistance from the therapist. Topics for both negotiation sessions were chosen by the couple before they began the discussions. Therapist instructions specified that the couple was to act as they would normally and to problem solve as best they could. These negotiation sessions were videotaped and then coded by trained observers using the Marital Interaction Coding System (MICS; Hops, Wills, Pat-

¹ The therapists were Bark Brasted, Katherine Engel, Kenneth Freedland, Debra Jackson-Spangler, and Jan Robertson.

terson, & Weiss, Note 2), a 29-category observational system that provides a sequential accounting of verbal and nonverbal communication processes. Validation of the MICS comes both through treatment studies, which demonstrated that the MICS was moderately sensitive in discriminating preintervention to post-intervention changes (Patterson, Hops, & Weiss, 1975), and analogous studies, which demonstrated that the MICS accurately discriminated couples who defined themselves as distressed or nondistressed (Birchler, Weiss, & Vincent, 1975). MICS coders were undergraduates, who were trained for a total of 20 hours or until they reached the minimum criterion of 70% reliability. Coders worked in pairs, and two pairs coded each 10-minute segment. To calculate percentage of agreement, one coder was randomly chosen to be "calibrator." The total frequency of agreements between coders was divided into the total number of codes recorded by the calibrator; that is, agreements plus disagreements. This quotient was then multiplied by 100. Any observation that did not reach the criterion of 70% reliability was recoded by another coder pair. High interobserver reliability was maintained by once per week random spot checks on individual coders.

Intervention

Communication training. The main therapeutic purpose of all three treatment conditions was to improve communication skills. The general structure used by all groups consisted of 10-minute negotiation discussions, during which spouses attempted to resolve a conflictual issue, and 10-minute feedback periods, during which spouses shared reactions to the process of the preceding negotiation discussion. Each 2-hour therapy session included four negotiation sessions interspersed with four feedback discussions. Specific procedures used during the negotiation and feedback discussions varied according to group assignment.

Communication training for the BT group incorporated principles common to many behavioral marital treatments, namely that spouses must (a) explicitly define (pinpoint) specific behaviors that the partner is to accelerate and then (b) faithfully reinforce the occurrence of these behaviors. The specific behaviors to be increased were "helpful" communications, which were operationally defined by the partitions. During each negotiation discussion, one spouse was designated as the "sender" and the other as the "receiver" of helpful communications. While attempting to resolve a problem issue, BT couples used an electromechanical apparatus to increase the rate at which the sender emitted helpful responses.

The electromechanical apparatus consisted of (a) a raucous buzzer functioning as a negatively reinforcing stimulus and (b) a pleasant single-tone chime functioning as a reward. The buzzer was activated by a recycling interval timer, which timed an interval of 1.75 minutes. The spouse designated as receiver of coded, by means of a silent hand-held button, each time the sender emitted a helpful response. Each cod-

ing of a response (a) activated the pleasant chime signaling to the sender that she/he had been helpful and (b) reset the timer to postpone the buzzer. If the 1.75 interval expired without a helpful response, the timer activated the buzzer, signaling to the sender that she/he must emit a response to terminate the buzzer. Spouses alternated roles for each new 10-minute negotiation discussion. During the feedback periods, the receiver for the previous discussion identified which behaviors the sender was to accelerate. The sender then rehearsed unfamiliar responses before his/her next trial as sender. In these BT procedures, each spouse identified what the partner could do to increase communication helpfulness.

In the AB condition, spouses also defined behaviors that they wanted accelerated. However, only those behaviors agreed on by both partners were accepted as targets for change. This strategy was adopted to avoid the problems that arise when one spouse demands a change that the other is reluctant to make. In a manner similar to BT couples, AB couples used the electromechanical apparatus and alternated between sender and receiver roles. The difference, however, was that AB couples worked to increase the rate at which they agreed on the coding of helpful responses rather than to increase one spouse's output of these responses (cf. Margolin & Weiss, in press). In the AB condition, both spouses simultaneously coded helpfulness. The sender coded helpful behaviors that she/he emitted, and the receiver coded the sender's behaviors that were perceived as helpful. The goal of the exercise was to increase sender-receiver agreements, which were defined as simultaneous responding by both partners within a 2-sec interval. Each of these agreements activated the chime and reset the buzzer. During feedback periods, AB spouses identified which communication behaviors they had jointly defined as helpful and then pinpointed and rehearsed these behaviors. Success on this task was labeled as couple agreement; that is, the spouses held congruent perceptions about the process of their communication even though they were still in conflict over content issues.

NS communication training followed a similar structural format of negotiation and feedback discussions but lacked specific behavioral training. During the negotiation discussions, spouses did not use the electromechanical cuing apparatus, nor were they in sender-receiver roles. Instead, they engaged in uninterrupted 10-minute discussions that focused on a relationship problem. During the feedback period, the therapist encouraged the spouses to express feeling statements and discuss perceptions but discouraged prolonged discharge of pent-up emotion. Rather than helping spouses to define specific behaviors, the therapist directed his/her activity toward the reflection and acceptance of each spouse's feelings.

Increasing the pleasing behaviors. The secondary therapeutic purpose across treatment conditions was to generate immediate gratifications for couples to offset the tedium of communication skill training. This treatment goal was achieved through homework assignments using the SOC. NS couples continued to

use the SOC in the same manner as all couples had done during the assessment phase. Each evening, NS spouses indicated which pleasing and displeasing items had occurred during the previous 24-hour period. In addition to tracking those events, each spouse in the BT group pinpointed desirable events on the SOC and worked toward a 100% increase in his/her output of those events that the partner had identified as pleasing. AB couples first agreed on a subset of events that they both identified as pleasing and then worked as a team toward a 100% increase in the identified couple pleasing behaviors.

Reading assignments. Assignment of the following readings lent rationale and credibility to the procedures for each group: NS couples read a chapter from Lederer and Jackson's (1968) *Mirages of Marriage* about communication in a marital relationship. BT couples read Chapter 2 from a manuscript by Weiss and Ford (Note 3) that defines behavioral principles, such as pinpointing, reinforcement, and shaping and demonstrates how these principles apply to marital interactions. AB couples read Chapters 1 and 2 of the Weiss and Ford manuscript; Chapter 1 describes how unsatisfying relationships are a function of poor relationship habits rather than either spouse's bad intentions.

Cognitive restructuring. In addition to the other therapeutic components, AB treatment included procedures to modify spouses' cognitions regarding the source of their marital discontent. The cognitive restructuring component, which was conveyed through therapist explanations and readings, contained the following messages: (a) Blame does not reside with one or the other spouse; (b) both persons suffer as participants in a dysfunctional relationship; and (c) frustration from feeling powerless to improve the relationship is often confused with resentment toward the spouse. Spouses were encouraged to apply non-blaming explanations to situations in which the partner's behavior was perceived as undesirable.

Results

Validity of Treatment

Two inventories completed at the conclusion of treatment provided an internal validity check on whether the treatment conditions captured the intent of the experimental design. A one-way analysis of variance of the Posttreatment Evaluation Questionnaire, which measured consumer satisfaction, revealed no group differences in how couples perceived or evaluated the different treatment modalities. The Consultation Readiness Inventory provided information on the dual questions of (a) whether couples in both behavioral groups were adequately instructed in social learning theory principles and (b)

whether, in fact, couples in the nonspecific group were privy to the same social learning theory information provided to the two behavioral groups. Findings on this inventory also confirmed design expectations. Mean item scores for behavioral (22.44) and behavioral-attitudinal (22.11) couples were significantly higher than the scores for nonspecific couples (12.22), $F(2, 24) = 11.06$, $p < .01$.

Outcome Criteria

Each outcome variable was analyzed for within-group and between-groups changes. Separate matched-pair t tests were used for within-group analyses to determine pretreatment to posttreatment changes for each group on each variable. Given the large number of analyses generated by this statistical approach, any particular test meeting the .05 level of statistical significance must be interpreted cautiously. The relative effectiveness of different treatment conditions was examined by between-groups analyses. Covariate procedures were used on all between-groups analyses to control for differences in prescores on each measure. Planned-comparison t tests on the adjusted posttreatment means were used when between-groups differences had been predicted; analysis of covariance was applied when there were no such predictions. In each statistical analysis, couples were analyzed as units; couples' scores were either the sum or the average of husband-wife scores.

Self-report satisfaction inventories. Table 1 presents the results for the four self-report measures. Within-group effects were measured by comparing pretreatment and posttreatment scores, obtained prior to therapist contact and after the posttreatment assessment week; t values are presented for pretreatment to posttreatment comparisons. Between-groups effects were examined through the adjusted posttreatment means.

The within-group marital satisfaction scores on the Locke-Wallace revealed that both the AB and the NS groups reported significantly increased satisfaction. In addition, significant between-groups differences were

Table 1
Treatment Effects on Self-Report Satisfaction Inventory Measures

Dependent variable	Within group				Pre-post <i>t</i>	Between-groups adjusted posttreatment <i>M</i>
	Pretreatment		Posttreatment			
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Locke-Wallace ^a						
BT	75.05	22.22	80.27	22.84	-1.10	76.81
AB	63.22	14.63	87.77	27.15	-3.99***	96.76 ^b
NS	77.05	21.38	84.16	24.16	-2.89**	78.55
Areas of change ^{a,c}						
BT	11.72	3.09	9.86	5.11	1.50	10.06
AB	13.44	5.19	7.22	4.93	3.25**	6.30
NS	11.00	3.67	9.66	4.93	1.34	10.37
ACL positive adjectives ^d						
BT	66.22	14.92	53.44	17.39	4.48 ^e	42.30
AB	45.55	20.40	45.88	23.49	-.26	54.74
NS	51.66	19.61	49.77	21.33	.48	52.06
ACL negative adjectives ^{a,d}						
BT	15.88	8.22	13.66	8.57	2.79*	14.55
AB	21.00	7.87	12.88	8.05	4.15***	9.97
NS	14.33	7.28	12.77	10.68	.41	11.41

Note. BT = behavioral training plus nonspecific effects; AB = attitudinal restructuring plus behavioral training and nonspecific effects; NS = nonspecific effects, ACL = Adjective Check List.

^a Average of husband-wife scores.

^b AB is significantly higher than BT and NS scores.

^c Low score = high marital accord.

^d Sum of husband-wife scores.

^e Change occurred in direction opposite to prediction.

* $p \leq .025$, one-tailed.

** $p \leq .01$, one-tailed.

*** $p \leq .005$, one-tailed.

found for adjusted posttreatment means, $F(2, 23) = 3.59$, $p \leq .05$; a post hoc Scheffé analysis on these data supported the conclusion that the AB group's posttreatment scores were significantly higher ($p < .05$) than those of the other two treatment conditions.

A similar pattern of results was obtained for the Areas of Change Questionnaire: Only the AB group significantly reduced total conflict scores from pretreatment to posttreatment measurements. Since all groups improved, between-groups differences were not significant.

The Adjective Check List results were examined by separate analyses of mean frequencies of checked positive and negative adjectives. The AB group significantly reduced their mean frequency of negative labels, although the mean frequency of positive

labels remained the same. The BT group reduced the frequency of both negative and positive adjectives, perhaps reflecting a tendency to use all trait labels less often. Couples in the NS group did not change from baseline rates on either adjective measure. No between-groups differences were found for mean frequencies of either positive or negative adjectives checked.

Overall, the AB group, relative to the other groups, consistently demonstrated a greater degree of improvement on the three self-report measures. Contrary to prediction, reported marital satisfaction as evidenced on the Locke-Wallace, increased in only two of the three groups.

Daily reported home observations. Spouses recorded daily pleasing and displeasing events received from the partner, that is, SOC items, as well as pleasant thoughts

Table 2

Treatment Effects on Measures of Mean Daily Recorded Home Data (Frequency of Items)

Dependent variable	Pretreatment		Within-group treatment		Posttreatment		Pre-post <i>t</i>	Between-groups adjusted posttreatment <i>M</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
SOC pleases								
BT	25.35	10.02	36.48	20.87	36.57	24.85	-1.45	36.48*
AB ^b	26.31	12.38	37.78	13.99	41.96	15.22	-4.74***	41.27*
NS	25.73	16.18	21.95	14.08	24.13	14.58	.55	24.87
SOC displeases								
BT	5.41	4.75	4.28	2.77	1.94	1.90	2.52*	2.25
AB	5.01	3.45	3.35	.81	3.29	2.28	2.54*	3.17
NS	5.67	4.64	3.90	2.38	2.10	2.12	3.25**	2.00
Pleasant thoughts								
BT	4.79	2.68	5.02	3.06	5.31	2.96	-1.17	4.74
AB	3.68	1.47	4.33	2.69	5.18	2.04	-2.37*	5.46
NS	3.77	4.16	3.97	3.76	4.06	4.36	-.38	4.13

Note. BT = behavioral training plus nonspecific effects; AB = attitudinal restructuring plus behavioral training and nonspecific effects; NS = nonspecific effects; SOC = Spouse Observation Checklist.

* BT and AB scores are significantly higher than NS scores.

^b Home observation data were missing for one AB couple since the husband was scheduled to leave for reserve duty immediately upon completing the 2-week treatment.

* $p \leq .025$, one-tailed.

** $p \leq .01$, one-tailed.

*** $p \leq .005$, one-tailed.

about that person. The AB and BT groups were expected to demonstrate greater improvements than the NS group, and the AB group was expected to display greater improvement than the BT group. Separate planned-comparison *t* tests were used to analyze these directional predictions.

Table 2 presents mean daily spouse pleasing, displeasing, and pleasant thought totals. Since the data were collected throughout treatment, the table also includes midtreatment scores averaged across the 14-day intervention. The pretreatment and posttreatment means are for 5 and 7 days, respectively, of data collection.² Pleasing events were explicitly increased as an intervention strategy in both behavioral groups thereby limiting the utility of recorded pleasing behaviors as an index of treatment outcome. This measure assesses (a) whether clients had followed the instructions to increase pleasing events during intervention and (b) whether these behavior changes were maintained after the instructions had been withdrawn.

During intervention both behaviorally oriented groups increased their mean daily

rates of pleases approximately 43% over their baseline rates. At posttreatment assessment, the AB group further increased its mean rate of pleasing behaviors, whereas the BT group merely maintained its intervention rate. Only the AB group demonstrated significant pretreatment to posttreatment change in rates of pleasing behaviors. Planned-comparison *t* tests confirmed the prediction that both behavioral groups would exchange more pleasing events than the NS group, $t(23) = 2.15$, $p \leq .025$. Examination of rates of displeasing behaviors revealed results unlike those for pleasing behaviors; the mean daily rate of displeasing behavior decreased significantly for all three groups.

The cognitive restructuring procedures were expected to produce significant increases in spouse-related pleasant thoughts. The AB group confirmed this expectation. Even though the other groups showed

² Data from the first 2 days of pretreatment were discarded because of their excessive variability for all groups.

changes indicating improvement, they were not significant.

Results of the daily recorded home data suggest that pleasing and displeasing behaviors were differentially effected by the treatment components. All groups displayed decreases in displeasing behaviors that were not even the target of any specific intervention technique. Yet increases in pleasing behaviors occurred only when there was a direct intervention to change those responses; merely tracking the occurrence/nonoccurrence of pleasing behaviors was not associated with changes in reported frequency of those behaviors.

Laboratory observational data. Communication skillfulness was assessed by MICS coding of two pretreatment and two post-treatment 10-minute videotaped samples of each couple negotiating a relationship problem. Average point-by-point observer agreement for the 108 coded samples was 83.8%. Coding was reduced to two summary scores for purposes of data analysis: positive responses (agree, approval, accept responsibility, compromise, humor, problem solution, attention, assent, laugh, positive physical contact, and smile) and negative responses (complaint, criticize, deny responsibility, excuse, no response, not tracking, put down, and turn off). Neutral behaviors such as

unfocused problem description were not included in the analyses. To control for differences in individual activity rates, each of these summary categories was expressed as the proportion of total responsiveness; individual husband and wife proportions were averaged for the couple score.

Table 3 provides pretreatment, posttreatment, and adjusted mean proportion scores for MICS coded positive and negative responses. Only the AB group significantly increased positive communication skills between pretreatment and posttreatment measurements. The mean increase in the positive response proportion was approximately 35% over baseline for the AB group, as compared to 9% and 20% increases for the BT and NS groups, respectively. A planned-comparison *t* test for between-groups measures revealed that the AB group increased significantly more than the BT group.

The between-groups differences found with the positive communication score were not paralleled with the negative communication score. Regardless of treatment mode, all groups reduced their mean proportion of negative behaviors to less than 50% of their baseline levels. The magnitude of these reductions was comparable for all treatment groups.

These results are consistent with the find-

Table 3
Treatment Effects on Laboratory Observational Data (Mean % Positive and Negative MICS Scores)

Dependent variable	Pretreatment		Posttreatment		Pre-Post <i>t</i>	Between-groups adjusted posttreatment <i>M</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Proportion positive						
BT	38.51	12.56	41.99	7.81	-.78	41.87
AB	35.29	8.44	47.71	5.38	-2.81**	47.66*
NS	34.01	11.73	40.83	6.67	-1.83	40.88
Proportion negative						
BT	9.33	10.34	3.75	4.18	1.86*	3.85
AB	8.80	6.60	3.49	2.82	2.13*	3.91
NS	11.30	6.30	5.50	4.38	4.21***	5.26

Note. BT = behavioral training plus nonspecific effects; AB = attitudinal restructuring plus behavioral training and nonspecific effects; NS = nonspecific effects.

* AB score is significantly higher than the BT score.

* $p \leq .05$, one-tailed.

** $p \leq .025$, one-tailed.

*** $p \leq .005$, one-tailed.

ings for SOC pleasing and displeasing behaviors; the between-groups difference for increases in desirable behaviors demonstrates strong support for the inclusion of the additional features of attitudinal restructuring. Reductions in negative responses do not appear to be related to specific therapeutic procedures, since all treatments were equally effective in producing a decrease in these behaviors.

Summary of Results

Overall, nine outcome measures spanning three modes of observation figured into the within-group and between-groups analyses. Out of the nine within-group *t* tests, the AB, BT, and NS groups showed significant changes in the predicted direction on 8, 3, and 3 measures, respectively. Significant between-groups differences were found on the following three outcome criteria: Locke-Wallace scores, SOC pleasing behaviors, and MICS positive behaviors. The AB group displayed significantly greater improvement than the BT group on the first and third measure and greater improvement than the NS group on the first measure. Thus, the AB group demonstrated significantly greater improvement than one or both of the other groups on at least one measure in each assessment mode.

Discussion

This study presented a method for implementing a therapeutic intervention for marital distress while maintaining careful control over the application of specific treatment procedures. Although a representative clinical sample was used, the following features of the study limit its applicability to naturalistic behavior marital therapy: (a) a highly abbreviated intervention period; (b) inexperienced paraprofessional therapists; and (c) highly standardized procedures. Because of the abbreviated nature of the intervention, the results of this study reflect changes that would occur during the initial stage of a standard course of therapy. It is as yet undetermined whether the demonstrated improvement would persist over time. Due to clinical considerations, follow-up data

were not collected to examine the maintenance of treatment effects. Since couples in this study received only a portion of a complete treatment, they were offered additional counseling through a marital treatment group. Follow-up data on the original study would have been confounded by this additional treatment, since only a portion of the couples participated in these groups and the participating couples joined the groups at different times after completing the original treatment.

The purpose of this study was to explore how specific therapeutic components differentially effect the initial stage of behavioral marital therapy. The results of this study revealed two predominant trends. Couples in the AB group, who received all three of the identified therapeutic components, consistently demonstrated greater improvement than couples in the other treatment groups. However, this finding applied only to criteria measuring change in a positive direction, that is, the measurement of attained relationship benefits. Criteria measuring dysfunctional coupling behaviors revealed significant reduction for all treatment groups; the treatment groups were not differentially effective in lessening problematic relationship behaviors.

There are several explanations to account for the reduction in negative behaviors displayed by all treatment groups. First, it is possible that demand characteristics associated with entering counseling heighten the level of negativity displayed at that time. Negative MICS behaviors would be high if spouses entered therapy with the notion that therapy is a time to argue and air grievances. The negative behaviors might then lessen over the weeks if the therapist did not reinforce these behaviors or if the spouses themselves grew tired of these unproductive patterns. Likewise, SOC displeasing behavior would be high if spouses used that instrument as a vehicle to demonstrate to the therapist the faults of the partner. SOC displeasing behavior might also decrease once the therapist communicated that she/he understood what each spouse was experiencing. Second, since the NS group received the same assessment procedures as the other

treatment groups, the reductions in negative behaviors displayed across all three groups may be indicative of clients' reactivity to the behavioral assessment procedures (cf. Johnson & Bolstad, 1973; Jones, Reid, & Patterson, 1975). That is, spouses' knowledge that their behaviors were being observed may have motivated them to reduce their negative behaviors in an effort to avoid censure by the partner or therapist. Unfortunately, reactivity to observations by the partner (on the SOC) or the therapist (through the MICS) cannot be distinguished from other nonspecific effects. It is interesting to note, however, that the data do not support an interpretation of similar reactivity in the positive direction; that is, spouses in the NS group did not increase positive or facilitative behaviors.

In light of more positive outcome literature on behavioral marital therapies (e.g., Jacobson, 1977; Weiss et al., 1973), what accounts for the BT group's relative ineffectiveness in this study? The brevity of the treatment may be one explanation; although the cognitive component appears to generate more rapid improvement, standard BT therapy might produce the same changes over a normal course of therapy. Second, the experimental design of this study excluded all cognitive restructuring factors from the BT group. Perhaps the systematic elimination of cognitive restructuring features in the BT treatment actually removed a competent that is typically included in behavior therapy under the rubric of reeducation or therapist persuasiveness. Most writings about behavioral treatments emphasize specific behavioral techniques and take for granted the procedures used to help clients generate a cognitive set that facilitates behavior change. Even though this study illustrates the pitfalls associated with behavioral training conducted in the absence of adequate cognitive preparation, more careful description of therapeutic procedures is needed to determine the extent to which behavior therapists currently use cognitive restructuring procedures.

Overall, the results of the present study suggest that the combination of the three identified therapeutic components in the AB

group provided an effective method to control aversive exchanges as well as an increase in rewarding marital interactions. These results pose the question as to which aspect of the cognitive restructuring component made the AB treatment more effective than the BT treatment. The two cognitive restructuring features that differentiated the AB from the BT treatment were (a) the modification of therapeutic exercises to focus on mutual rather than individual goals and (b) the reattribution message conveyed through readings and therapist instructions. Anecdotal feedback from clients suggested that the focus on mutual goals was helpful for distressed spouses who were reluctant to engage in behavior changes requested by the partner. The mutual goal setting allowed each spouse to exercise more control in defining the behaviors that she/he was to change. However, for spouses who share very few common goals, the mutual focus may limit the scope of therapeutic change and impede overall progress.

Cognitive restructuring within a marital therapy framework provides the most innovative aspect of the AB treatment. Recently, a number of investigators working on a variety of problem areas (cf. Mahoney, 1974; Meichenbaum, 1977) have used cognitive restructuring procedures to reduce the incapacitating emotions that tend to accompany dysfunctional behaviors, making those behaviors resistant to change. The cognitive component in this study was an initial attempt in using cognitive restructuring procedures to interrupt the negative thoughts that can impede relationship improvement. The therapeutic procedures encouraged spouses to abandon blaming attributions, to accept greater personal responsibility for relationship failure, and to be more accepting of their partners' positive efforts. This approach, which is somewhat similar to Ellis' (1962) rational-emotive psychotherapy, was expected to decrease the likelihood of mislabeling ambiguous behaviors or overresponding to negative relationship exchanges. The reattribution procedures in this study were presented in a standardized fashion that disallowed their being tailor-made to a couple's particular needs. Furthermore, no attempt was made to

rehearse the cognitive restructuring materials or to monitor how spouses used them. Nonetheless, the encouraging findings for the AB treatment strongly support the application of cognitive restructuring procedures to reduce spouses' blaming attributions and to foster prerequisite attitudes for constructive problem solving.

Reference Notes

1. Turkewitz, H., & O'Leary, K. D. *A comparison of communication and behavioral marital therapy*. Paper presented at the annual meeting of the Association for the Advancement of Behavior Therapy, Atlanta, December 1977.
2. Hops, H., Wills, T. A., Patterson, G. R., & Weiss, R. L. *Marital Interaction Coding System*. Unpublished manuscript, University of Oregon and Oregon Research Institute, 1972.
3. Weiss, R. L., & Ford, L. *A social learning view of marriage*. Unpublished manuscript, University of Oregon, 1975.

References

- Birchler, G. R., Weiss, R. L., & Vincent, J. P. A multimethod analysis of social reinforcement exchange between maritally distressed and nondistressed spouse and stranger dyads. *Journal of Personality and Social Psychology*, 1975, 31, 349-360.
- Ellis, A. *Reason and emotion in psychotherapy*. New York: Lyle Stuart, 1962.
- Gough, H. G., & Heilbrun, A. B., Jr. *The Adjective Check List manual*. Palo Alto, Calif.: Consulting Psychologists Press, 1965.
- Gurman, A. S., & Kniskern, D. P. Research on marital and family therapy: Progress, perspective, and prospect. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (2nd ed.). New York: Wiley, in press.
- Harrell, J., & Guernsey, B. Training married couples in conflict negotiation skills. In D. H. L. Olson (Ed.), *Treating relationships*. Lake Mills, Iowa: Graphic Press, 1976.
- Jacobson, N. S. Problem-solving and contingency contracting in the treatment of marital discord. *Journal of Consulting and Clinical Psychology*, 1977, 45, 92-100.
- Jacobson, N. S. The effects of problem-solving on increasing positive marital behavior. *Behavior Therapy*, in press.
- Jacobson, N. S. A review of the research on the effectiveness of marital therapy. In T. Paolino & B. McCrady (Eds.), *Marriage and the treatment of marital disorders from three perspectives: Psychoanalytic, behavioral, and systems theory*. New York: Bruner/Mazel, 1978. (a)
- Jacobson, N. S. Specific and nonspecific factors in a behavioral approach to marital discord. *Journal of Consulting and Clinical Psychology*, 1978, 46, 442-452. (b)
- Jacobson, N. S., & Martin, B. Behavioral marriage therapy. *Psychological Bulletin*, 1976, 83, 540-556.
- Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In F. W. Clark & L. A. Hamerlynck (Eds.), *Critical issues in research and practice: Proceedings of the Fourth Banff International Conference on Behavior Modification*. Champaign, Ill.: Research Press, 1973.
- Jones, R. R., Reid, J. B., & Patterson, G. R. Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 3). San Francisco: Jossey-Bass, 1975.
- Lederer, W. J., & Jackson, D. D. *The mirages of marriage*. New York: Norton, 1968.
- Liberman, R., Levine, J., Wheeler, E., Sanders, N., & Wallace, C. Marital therapy in groups: A comparative evaluation of behavioral and interactional formats. *Acta Psychiatrica Scandinavica Monograph*, 1976, Supplement 226, 1-34.
- Locke, H. J., & Wallace, K. M. Short marital adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living*, 1959, 21, 251-255.
- Mahoney, M. J. *Cognition and behavior modification*. Cambridge, Mass.: Ballinger, 1974.
- Margolin, G., Christensen, A., & Weiss, R. L. Contracts, cognition, and change: A behavioral approach to marriage therapy. *Counseling Psychologist*, 1975, 5, 15-26.
- Margolin, G., & Weiss, R. L. Communication training and assessment: A case of behavioral marital enrichment. *Behavior Therapy*, in press.
- Meichenbaum, D. *Cognitive-behavior modification: An integrative approach*. New York: Plenum Press, 1977.
- Olson, D. H., & Ryder, R. G. Inventory of Marital Conflicts (IMC): An experimental interaction procedure. *Journal of Marriage and the Family*, 1970, 32, 443-448.
- Patterson, G. R., Hops, H., & Weiss, R. L. Interpersonal skills training for couples in early stages of conflict. *Journal of Marriage and the Family*, 1975, 37, 295-302.
- Weiss, R. L., Hops, H., & Patterson, G. R. A framework for conceptualizing marital conflict, a technology for altering it, some data for evaluating it. In F. W. Clark & L. A. Hamerlynck (Eds.), *Critical issues in research and practice: Proceedings of the fourth Banff International Conference on Behavior Modification*. Champaign, Ill.: Research Press, 1973.
- Weiss, R. L., & Margolin, G. Marital conflict and accord. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook for behavioral assessment*. New York: Wiley, 1977.

Comments

Bias or Artifact? A Reply to Reade and Wertheimer

David J. McDowell

University of Maine and Worcester State Hospital, Worcester, Massachusetts

In a recent *Journal of Consulting and Clinical Psychology* brief report, Reade and Wertheimer reported an analogue study of the diagnosis of schizophrenia and inferred from their data that diagnosticians' judgments of the probability of schizophrenia are biased inappropriately by the information that a patient has an identical twin diagnosed as schizophrenic. In rebuttal, it is argued that the procedures used in that study compelled the diagnostic behavior that the authors criticized and that the implicit generalization from this analogue study to reasonable clinical practice was unjustified.

Reade and Wertheimer (1976) reported an analogue diagnostic study in which a brief case description was given to two groups of clinical judges who rated the probability that the fictitious patient was schizophrenic. The two case descriptions differed only in their second sentence, which stated either that "Mr. S has an identical twin, who 7 years ago joined the army and has visited home only once for a 2-week long stay" (p. 878), or "Mr. S has an identical twin, who 4 years ago was placed in a mental hospital with a diagnosis of simple schizophrenia" (p. 878). The remainder of the case study described parental roles (father was domineering and quick-tempered; mother was attentive and spoiling), emphasized the patient's poor interpersonal relations (introverted, shy with girls, few close relationships), and referred to some possibly symptomatic behaviors (refusing to speak to his father for several days after his father berated him for not looking for a job, apathy, and "a slow but pronounced change in his character," which included untidiness, neglect of personal appearance, "strange" behavior at his father's funeral, and frequent change of jobs due to poor peer relations and malingering). The clinicians who read that Mr. S had a schizophrenic identical twin rated a mean probability of schizophrenia of .66, whereas the clinicians who read that Mr. S's twin

had joined the army rated a mean probability of schizophrenia of .39. The two groups' ratings differed significantly ($p < .01$) by a Mann-Whitney U test, and the authors concluded that the information that Mr. S had a twin diagnosed as schizophrenic was responsible for this difference.

Should such information increase the clinician's likelihood of diagnosing a patient as schizophrenic? Reade and Wertheimer (1976) argued that diagnosis "should ideally be based on an analysis of individuals' . . . behavior . . . [and not on the] knowledge that a predisposition to schizophrenia may be inherited." They concluded that

prior to availing themselves of family history, diagnosticians [should] attempt to obtain relatively conclusive behavioral information, lest they inadvertently increase the likelihood of a diagnosis of psychopathology where none exists" (p. 878).

Although this argument and conclusion are sound, neither is relevant to the data reported.

In this study, clinicians (a) were not permitted to observe the patient's behavior; (b) were given pertinent family history before, and instead of, good behavioral records; and (c) were asked to indicate the probability of a specific diagnosis rather than make their own diagnosis in a free-response format. The procedures used in the study thus compelled the clinical-diagnostic behavior that the investigators then decry!

What the reported data *do* indicate is that in the absence of direct observation or good be-

Requests for reprints should be sent to David J. McDowell, who is now at Milford Assistance Program, Inc., P.O. Box 365, Milford, Massachusetts 01757.

havioral records, and with an extremely restrictive set of data from which to assess the probability of a specified diagnosis, these clinicians attended to a relevant family history variable that has been shown to correlate significantly with the diagnosis in question. Reade and Wertheimer implied that the diagnostic judgment of these clinicians was biased by nonbehavioral data. On the contrary, their experiment demonstrates (a) that clinicians are aware of the significant concordance rates for schizophrenic identical twins and their undiagnosed probands and (b) that

the analogue study of diagnostic judgment may differ sufficiently from reasonable clinical practice that generalization to the latter is unjustified.

Reference

- Reade, W. K., & Wertheimer, M. A bias in the diagnosis of schizophrenia. *Journal of Consulting and Clinical Psychology*, 1976, 44, 878.

Received April 11, 1977 ■

Psychometric Phrenology Revisited: Comments on Neuropsychological Testing

Steven Waksman

Multnomah County School Mental Health Program
Portland, Oregon

Amante, Van Houten, Grieve, Bader, and Margules' interpretation of results obtained from testing third-grade children with the Bender-Gestalt Test, the Frostig Developmental Test of Visual Perception, and the Wepman Auditory Discrimination Test is critically evaluated. It is pointed out that such tests not only lack empirical and linguistic validity, but they also tend to confuse their users and often hamper more appropriate remediation services to children with learning disabilities. The continued support of such tests is questioned.

In a recent article in this journal, Amante, Van Houten, Grieve, Bader, and Margules (1977) described the racial and socioeconomic differences obtained on commonly administered tests of perceptual functioning. Although the article raises some important and interesting issues concerning environmental factors related to neuropsychological functioning, it unfortunately stumbles into the logical abyss occupied by much of the literature concerning learning disabilities and neurological deficits.

The authors tested

a random sample of 225 third-grade public and parochial school children of both sexes representing a cross-section of subjects drawn from all of the major ethnic groups and socioeconomic levels" (Amante et al., 1977, p. 525)

present in their community. The results indicated that between 40%-60% of the sample exhibited "evidence" of perceptual-motor, visual-perception, and auditory discrimination handicaps as measured by the Bender-Gestalt Test, the Frostig Developmental Test of Visual Perception, and the Wepman Auditory Discrimination Test, respectively. The authors then proceeded to interpret such results as indicating neuropsychological deficits and discussed the implications of the ethnic and socioeconomic differences.

I wonder how psychologists and educators can continue to support the use of these tests for diagnosing neuropsychological problems when

they identify 40%-60% of randomly selected children as neuropsychologically handicapped or impaired. Undoubtedly, the vast majority of these "handicapped" children were functioning appropriately or excelling both academically and socially in their environments. The lack of linguistic (Payne, 1975) and empirical (Anderson, 1965; Blakemore, 1965; DiCarlo, 1965; Goldberg, 1959; Larson, Rogers, & Sowell, 1976) validity of such widely used instruments is appalling. As Jastak and Jastak (1976) have pointed out, not only are such tests poorly conceived and validated, but in reality they "convey no meaningful information" and only tend to confuse their users as well as the teachers and parents of the children on whom they are used.

This process is witnessed all too often by school personnel who refer students with learning problems for evaluation and recommendations. The evaluation reports that often return with the referred student describe perceptual deficits or handicaps as the "cause" of the learning problem. According to the Amante et al. (1977) data, every third-grade child has a 40%-60% chance of being diagnosed as perceptually handicapped. To confound the issue even more, there is a myriad of perceptual training programs available that offer no research support for their effectiveness (McIntosh & Dunn, 1973) and that often stand in the way of appropriate remediation attempts (Mann, 1970). In short, as Mann (1971) has previously argued, such "psychometric phrenology" and the perceptual training programs that often follow such ability assessment offer little more than the "philosophies of faculty psychology."

Requests for reprints should be sent to Steven Waksman, 2242 Northeast 21st Avenue, Portland, Oregon 97212.

Although perceptual problems are often a correlate of neurological impairment (Chalfant & Scheffelin, 1969; McIntosh & Dunn, 1973), the converse of this is not necessarily true. Learning problems may result from many factors (Eichenwald & Fry, 1969; Jastak & Jastak, 1976; McIntosh & Dunn, 1973; Moyer & Newcomer, 1977), and there is still no evidence as yet (Goodman, 1977) that neurologically impaired children (except in very rare cases) should be taught any differently from nonimpaired children.

The time has come for us to stop blaming children for their lack of academic achievement, to stop scaring parents with brain-damaged labels, and to start sweeping the field of learning disabilities clear of faulty measurement and foggy logic.

References

- Amante, D., Van Houten, V. W., Grieve, J. H., Bader, C. A., & Margules, P. H. Neuropsychological deficit, ethnicity, and socioeconomic status. *Journal of Consulting and Clinical Psychology*, 1977, 45, 524-535.
- Anderson, J. M. Marianne Frostig Developmental Test of Visual Perception (3rd ed.). In O. K. Buros (Ed.), *The sixth mental measurement yearbook*. Highland Park, N.J.: Gryphon Press, 1965.
- Blakemore, C. B. Bender-Gestalt Test. In O. K. Buros (Ed.), *The sixth mental measurement yearbook*. Highland Park, N.J.: Gryphon Press, 1965.
- Chalfant, J. C., & Scheffelin, M. A. (Eds.). *Central processing dysfunctions in children: A review of research, Phase III* (HINDS Monograph No. 9). Washington, D.C.: U.S. Government Printing Office, 1969.
- DiCarlo, L. M. Auditory Discrimination Test. In O. K. Buros (Ed.), *The sixth mental measurement yearbook*. Highland Park, N.J.: Gryphon Press, 1965.
- Eichenwald, H. F., & Fry, P. C. Nutrition and learning. *Science*, 1969, 163, 644-648.
- Goldberg, L. The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender-Gestalt Test. *Journal of Consulting Psychology*, 1959, 23, 25-33.
- Goodman, J. F. The diagnostic fallacy: A critique of Jane Mercer's concept of mental retardation. *Journal of School Psychology*, 1977, 15, 197-206.
- Jastak, J. F., & Jastak, S. R. *The Wide Range Achievement Test: Manual of instruction*. Wilmington, Del.: Guidance Associates of Delaware, 1976.
- Larson, S. C., Rogers, D., & Sowell, V. The use of selected perceptual tests in differentiating between normal and learning disabled children. *Journal of Learning Disabilities*, 1976, 9, 85-90.
- Mann, L. Perceptual training: Misdirections and re-directions. *American Journal of Orthopsychiatry*, 1970, 40, 30-38.
- Mann, L. Psychometric phrenology and the new faculty psychology: The case against ability assessment and training. *Journal of Special Education*, 1971, 5, 3-14.
- McIntosh, D. K., & Dunn, L. M. Children with major specific learning disabilities. In L. M. Dunn (Ed.), *Exceptional children in the schools*. New York: Holt, Rinehart & Winston, Inc., 1973.
- Moyer, S. B., & Newcomer, P. L. Reversals in reading: Diagnosis and remediation. *Exceptional Children*, 1977, 43, 424-429.
- Payne, J. L. *Principles of social science measurement*. College Station, Tex.: Lytton, 1975.

Received November 28, 1977 ■

Response to Psychometric Phrenology Revisited: Comments on Neuropsychological Testing

Dominic Amante

West Shore Mental Health Clinic, Muskegon, Michigan

Steven Waksman has raised some controversial issues in his critique of a neuropsychological study. The study dealt with the ecological distribution of central nervous system pathology in working class and socially disadvantaged black children and white children. Some of the issues that Waksman raises can be resolved, at least partially, but others cannot because of the fact that basic differences of opinion continue to exist in the absence of uncontested theory and decisive empirical data.

Waksman (1978) has raised some vital issues in his critique of a neuropsychological field study that we reported in this journal recently (Amante, Van Houten, Grieve, Bader, & Margules, 1977). Unfortunately, he has also introduced a variety of irrelevant issues.

It is important to observe that Waksman's critique is highly selective. For example, it does not deal with the intelligence test results. Consequently, it fails to refer to perhaps the most positive feature of the Amante et al. study—namely, the possibility that there are no differences in the levels of general intelligence or auditory discrimination in black children and white children of comparable socioeconomic status when the level of visual-motor functioning is controlled. His interest is in deficit conditions only, but even in this case he misrepresents our data and overgeneralizes the conclusions.

Based on an extensive clinical and research tradition, many medical and behavioral scientists have concluded that various forms of mild or severe neuropathology measurably influence a broad spectrum of psychological functions including language development, affective states, perceptual-motor skills, cognitive abilities, and various other behavioral parameters (Dimond & Beaumont, 1974; Reitan, 1975; Strub & Black, 1977). Consequently, there is no essential reason to conclude, as does Waksman, that the majority of perceptually impaired children were adequately functioning in their environments. Per-

formance on the Bender-Gestalt, for example, does relate to nontest factors such as school readiness, intellectual level, academic achievement, emotional and behavioral problems, and many diverse forms of neurological pathology (Koppitz, 1975).

It is doubtful if the test instruments that were used are as psychometrically unsophisticated as Waksman and others would have us believe. In fact, such techniques may be preferable considering the reliability and validity of simple behavioral observation, subjective impression, or "clinical intuition" (Korchin, 1976; Landy & Trumbo, 1976). It is true that modern medicine has been unable to devise neurodiagnostic techniques of perfect reliability and validity (Kennedy & Ramirez, 1964), and as nearly everyone now realizes, psychologists and educators have no instruments that are diagnostically impeccable (Anastasi, 1968). There is, however, a plausible rationale behind neuropsychological assessment (Boll, 1977), but in all cases we necessarily deal with diagnostic probabilities—not certitudes (Amante, 1976).

Some other issues should be considered. Waksman's (1978) primary interest appears to involve learning disabilities and related problems as well as tertiary prevention, which Caplan (1964) defines as an attempt to reduce the residual defect(s) associated with various conditions of physical and behavioral pathology. However, Amante et al. (1977) dealt with more general issues involving a diversity of neuropsychological outcomes, and the central thrust of the article pertains to primary and secondary prevention. That is, our basic concern involved the identification of environmentally based etiological factors, the control of which may reduce

Requests for reprints should be sent to Dominic Amante, West Shore Mental Health Clinic, 2525 Hall Road, Muskegon, Michigan 49442.

the incidence and the prevalence of central nervous system pathology in children.

It is conceded that the differential diagnosis and treatment of learning disorders and related problems is a controversial area (Larsen & Hammill, 1975; Ross, 1976; Spache, 1976). The interesting point that Waksman (1978) overlooks, however, is that it is also true that many of the leading theoreticians and researchers in the field believe that most (if not all) learning-disabled children are neurologically impaired, that they frequently do appear to have perceptual-motor and behavioral problems, and that various specialized educational methods individualized for the child often are called for (Cantwell, 1975; Johnson & Myklebust, 1967; Kenny & Clemmens, 1971; Kephart, 1969; Myers & Hammill, 1976; Reitan & Boll, 1973). William M. Cruickshank's (1972) position is not atypical:

I would like to suggest . . . that irrespective of the presence or absence of diagnosed neurological dysfunction, learning disabilities are essentially and almost always the result of perceptual problems based on the neurological system (p. 383).

Finally, it should be recognized that Waksman (1978) seems to assume the empirical viability of labeling theory (Scheff, 1975) even though many sociologists, who created the theory, have seriously questioned the validity of the approach (Gove, 1975). I agree that it is imperative from an ethical and a diagnostic point of view to exercise extreme caution when classifying children (Hobbs, Egerton, & Matheny, 1975), but if actual problems appear to be present, the diagnostician has an obligation to confront the situation and to devise a rational intervention strategy. We are not helpful to children if we evade our diagnostic responsibilities with the vague claim that somehow the diagnostic label creates more problems than it resolves. Some problems actually do exist in the real world, and if we fail to detect them, some children will be denied services that could be helpful to them, and preventive efforts in general will be seriously undermined.

References

- Amante, D. A. A neuropsychodiagnostic model. *International Journal of Neuroscience*, 1976, 6, 289-295.
- Amante, D., Van Houten, V. W., Grieve, J. H., Bader, C. A., & Margules, P. H. Neuropsychological deficit, ethnicity, and socioeconomic status. *Journal of Consulting and Clinical Psychology*, 1977, 45, 524-535.
- Anastasi, A. *Psychological testing*. New York: MacMillan, 1968.
- Boll, T. J. A rationale for neuropsychological evaluation. *Professional Psychology*, 1977, 8, 64-71.
- Cantwell, D. P. (Ed.). *The hyperactive child*. New York: Spectrum, 1975.
- Caplan, G. *Principles of preventive psychiatry*. New York: Basic Books, 1964.
- Cruickshank, W. M. Some issues facing the field of learning disabilities. *Journal of Learning Disabilities*, 1972, 5, 380-388.
- Dimond, S. J., & Beaumont, J. G. (Eds.). *Hemispheric function in the human brain*. New York: Wiley, 1974.
- Gove, W. R. (Ed.). *The labelling of deviance*. New York: Wiley, 1975.
- Hobbs, N., Egerton, J., & Matheny, M. H. Classifying children. *Children Today*, 1975, 4, 21-25.
- Johnson, D. J., & Myklebust, H. R. *Learning disabilities*. New York: Grune & Stratton, 1967.
- Kennedy, C., & Ramirez, L. S. Brain damage as a cause of behavior disturbance in children. In H. G. Birch (Eds.), *Brain damage in children*. Baltimore: Williams & Wilkins, 1964.
- Kenny, T. J., & Clemmens, R. L. Medical and psychological correlates in children with learning disabilities. *Journal of Pediatrics*, 1971, 78, 273-277.
- Kephart, N. C. Teaching the child with a perceptual motor handicap. In M. Bortner (Ed.), *Evaluation and education of children with brain damage*. Springfield, Ill.: Charles C Thomas, 1969.
- Koppitz, E. *The Bender Gestalt Test for young children* (Vol. 2). New York: Grune & Stratton, 1975.
- Korchin, S. J. *Modern clinical psychology*. New York: Basic Books, 1976.
- Landy, F. J., & Trumbo, D. A. *Psychology of work behavior*. Homewood, Ill.: Dorsey Press, 1976.
- Larson, S. C., & Hammill, D. D. The relationship of selected visual-motor perceptual abilities to school learning. *Journal of Special Education*, 1975, 9, 281-291.
- Myers, P. I., & Hammill, D. D. *Methods for learning disorders*. New York: Wiley, 1976.
- Reitan, R. M. Assessment of brain-behavior relationships. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol 3). San Francisco: Jossey-Bass, 1975.
- Reitan, R. M., & Boll, T. J. Neuropsychological correlates of minimal brain dysfunction. *Annals of New York Academy of Sciences*, 1973, 205, 65-88.
- Ross, A. O. *Psychological aspects of learning disabilities and reading disorders*. New York: McGraw-Hill, 1976.
- Scheff, T. J. (Ed.). *Labeling madness*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.
- Spache, G. E. *Investigating the issues of reading disabilities*. Boston: Allyn & Bacon, 1976.
- Strub, R. L., & Black, F. W. *The mental status examination in neurology*. Philadelphia: Davis, 1977.
- Waksman, S. Psychometric phenology revisited: Comments on neuropsychological testing. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1489-1490.

Can Clients Provide Valuable Feedback to Clinicians About Their Personality Interpretations? A Reply to Greene

C. R. Snyder, Mitchell M. Handelsman, and Janet R. Endelman
University of Kansas

Greene's article on the Barnum effect is criticized on conceptual, methodological, and statistical grounds. A reanalysis of the data is presented, along with background research that conflicts with Greene's conclusions regarding (a) the ability of people to judge the accuracy of personality interpretations that they receive and (b) the role of feedback uniqueness in that judgment.

In commenting on previous Barnum effect research, Greene (1977) suggested that the recipient of personality feedback "accepts an interpretation as being descriptive if it is accurate, regardless of its uniqueness" (1965, emphasis added). Contrary to this assertion, however, previous studies have indicated that the uniqueness of feedback does influence acceptance on two levels. First, a consistent finding is that the higher the base-rate accuracy for people in general (i.e., the lower the uniqueness), the greater the acceptance of the feedback (e.g., Collins, Dmitruk, & Ranney, 1977; Merrens & Richards, 1970; Mosher, 1965; O'Dell, 1972; Snyder & Shenkel, 1976; Sundberg, 1955; Weisberg, 1970). That is, the content of the interpretive feedback should not be unique in order to elicit acceptance on the part of the recipient. Second, it is important for the recipient of diagnostic feedback to believe that the feedback is derived specifically for him or her (i.e., that it is unique). Research on this point has shown that recipients of feedback are more acceptant of the feedback when they believe that it was "specifically derived for them" on the basis of psychological tests, as compared to "statements that are generally true of people" (Snyder, 1974; Snyder, Larsen, & Bloom, 1976; Snyder & Larson, 1972; Snyder & Shenkel, 1976; Ziv & Nevenhaus, 1972).

Other comments are warranted on the methodology used in the Greene (1977) study.

Greene used a fairly typical procedure in the Barnum effect paradigm (cf. Snyder, Shenkel, & Lowery, 1977, for review). Undergraduate students in two classes completed a psychological test in one class session, and in the next class session each student was given the same interpretive feedback. Students then rated the interpretation

according to the extent to which this personality interpretation described their own personality. Second, they rated the extent to which it described them as a unique individual, that is, as different from their classmates. Third, they rated the extent to which it described one of their classmates. (Greene, 1977, p. 965)

Subjects responded to each question on a 5-point scale (5 = excellent, 4 = good, 3 = average, 2 = poor, 1 = very poor). Methodologically, the within-subject sequence may have sensitized subjects to become highly skeptical. It would have been more desirable to counterbalance the order of items to handle this possibility. Also, it should be noted that any within-subject differences may be due in part to shifting internal anchor points resulting from different content across questions.

Given the aforementioned design, a Class (senior vs. junior-sophomore) \times Questions (1, 2, 3) analysis of variance would have been preferable to the single one-way analysis of variance that was reported. (In this subsequent analysis the reader should be cautioned not to overinterpret the results because of the previously mentioned problems with the three questions.) When Greene's (1977) data were reanalyzed this way, the main effects of class and questions were significant. The main effects were qualified by a significant Class \times Questions interaction, $F(2, 98) = 3.22, p < .05$. This inter-

The authors gratefully acknowledge the assistance of Roger Greene in providing the data of his study for further analysis.

Requests for reprints should be sent to C. R. Snyder, Graduate Training Program in Clinical Psychology, Fraser Hall, University of Kansas, Lawrence, Kansas 66045.

action resulted because the senior and junior-sophomore classes did not differ significantly on the "describe classmates" question, but the junior-sophomore class (a) accepted the interpretation more highly, $t(49) = 2.06$, $p < .045$, and (b) perceived the interpretation as being more unique for them than did the senior class, $t(49) = 2.90$, $p < .006$. Whether these differences between classes are due to greater sophistication on the part of the senior as compared to junior-sophomore class members, as Greene (1977) suggested, is not discernable from these data. This follows because the two classes may have differed on a multitude of possible dimensions that may influence their reactions to feedback (e.g., time of course in terms of day and semester, course content, instructor, or class composition in terms of sex, age, race, etc.). Also the inference made by Greene (1977) that students could realize "that the same interpretation could as accurately be applied to *any* of their classmates" (Greene, 1977, pp. 965-966, emphasis added) is overstating the results, since the actual statement read "rate the extent to which it describes *one* of your classmates." One other suggestion regarding the present procedure is that performing the experiment in a classroom setting may have aroused far more suspicion on the part of subjects than the individual testing and feedback situation that is typical of actual clinical practice and recent research within this paradigm. The literature suggests that the situation in which a client receives interpretive feedback from a clinician may yield a rather highly acceptant and undiscerning recipient of feedback (Snyder & Clair, 1977).

Greene's conclusion that "the Barnum effect can become a less formidable adversary if the clinician is careful to instruct the student or client as to exactly what questions are to be answered" (Greene, 1977, p. 966) is unwarranted given the previous mentioned methodological limitations. Additionally, although clinicians in actual practice may ask clients to give their acceptance reactions to feedback, they probably do not ask clients about the uniqueness of the feedback. Finally, it is unfortunate that the previous Barnum effect literature does not support Greene's other important conclusion, that "the student or client can be a source of valuable feedback to the clinician in honing his or her clinical skills" (Greene, 1977, p. 966). It has been found that students cannot distinguish bona fide personality interpretations from randomly generated or bogus interpretations (Dies, 1972; Sundberg, 1955). In fact, several studies even found that the generalized inter-

pretations are seen as more accurate than bona fide ones (Merrens & Richards, 1970; O'Dell, 1972).

An old quote comes to mind in regard to this entire issue. Upon hearing of his death, Mark Twain wrote the following cable to the Associated Press in 1897, "The reports of my death are greatly exaggerated" (cited in Bartlett, 1968, p. 763). Given the present state of research, the same could be said of the Barnum effect.

References

- Bartlett, J. *Familiar quotations* (14th ed.). Boston: Little, Brown, 1968.
- Collins, R. W., Dmitruk, V. M., & Ranney, J. T. Personal validation: Some empirical and ethical considerations. *Journal of Consulting and Clinical Psychology*, 1977, 45, 70-77.
- Dies, R. R. Personal gullibility or pseudodiagnosis? A further test of the "fallacy of personal validation." *Journal of Clinical Psychology*, 1972, 28, 47-50.
- Greene, R. L. Student acceptance of generalized personality interpretations: A reexamination. *Journal of Consulting and Clinical Psychology*, 1977, 45, 965-966.
- Merrens, M. R., & Richards, W. S. Acceptance of generalized versus "bona fide" personality interpretations. *Psychological Reports*, 1970, 27, 691-694.
- Mosher, D. L. Approval motive and acceptance of "fake" personality test interpretations which differ in favorability. *Psychological Reports*, 1965, 17, 395-402.
- O'Dell, J. W. P. T. Barnum explores the computer. *Journal of Consulting and Clinical Psychology*, 1972, 38, 270-273.
- Snyder, C. R. Acceptance of personality interpretations as a function of assessment procedures. *Journal of Consulting and Clinical Psychology*, 1974, 42, 150.
- Snyder, C. R., & Clair, M. S. Does insecurity breed acceptance?: Effects of trait and situational insecurity on acceptance of positive and negative diagnostic feedback. *Journal of Consulting and Clinical Psychology*, 1977, 45, 843-850.
- Snyder, C. R., Larsen, D., & Bloom, L. J. Acceptance of personality interpretations prior to and after receiving diagnostic feedback supposedly based on psychological, graphological, and astrological assessment procedures. *Journal of Clinical Psychology*, 1976, 32, 258-265.
- Snyder, C. R., & Larson, G. R. A further look at student acceptance of general personality interpretations. *Journal of Consulting and Clinical Psychology*, 1972, 38, 384-388.
- Snyder, C. R., & Shenkel, R. J. Effects of "favorability," modality, and relevance upon acceptance

- of general personality interpretations prior to and after receiving diagnostic feedback. *Journal of Consulting and Clinical Psychology*, 1976, 44, 34-41.
- Snyder, C. R., Shenkel, R. J., & Lowery, C. R. Acceptance of personality interpretations: The "Barnum effect" and beyond. *Journal of Consulting and Clinical Psychology*, 1977, 45, 104-114.
- Sundberg, N. D. The acceptability of "fake" versus "bona fide" personality test interpretations. *Journal of Abnormal and Social Psychology*, 1955, 50, 145-147.
- Weisberg, P. Student acceptance of bogus personality interpretations differing in level of social desirability. *Psychological Reports*, 1970, 27, 743-746.
- Ziv, A., & Nevenhaus, S. Acceptance of personality diagnosis and perceived uniqueness. *Abstract Guide of XXth International Congress of Psychology*, 1972, 605.

Received December 27, 1977 ■

Can Clients Provide Valuable Feedback to Clinicians About Their Personality Interpretations? Greene Replies

Roger L. Greene
Texas Tech University

Snyder, Handelsman, and Endelman's response to Greene overlooked several significant issues that are important in reconceptualizing research on students' acceptance of generalized personality interpretations. These issues are stated explicitly, and their relevance to research in this area is described.

Snyder, Handelsman, and Endelman (1978) have provided an interesting commentary on Greene's (1977) article on student acceptance of generalized personality interpretations. Somehow they have managed to overlook the major thesis of his article, and instead they have chosen to comment on several tangential issues. Since they overlooked the major point of his article, it will be reiterated here so that other researchers will not make the same oversight. Generalized personality statements are constructed so that these statements are an *accurate* (true) description of all students. Hence, it does not make sense then to ask students whether these statements are accurate when in fact they are constructed to be so. Unfortunately, most previous research on students' acceptance of generalized personality interpretations has asked the student how accurately such statements fit his/her personality (e.g., Forer, 1949; Snyder, 1974; Ulrich, Stachnik, & Stainton, 1963). The issue of importance is whether students believe these generalized personality interpretations to be unique.

Since Snyder et al. (1978) use the term *uniqueness* in two different senses, it is not clear exactly what point they are trying to make in their comment on this issue. Uniqueness can be used to describe the base-rate accuracy of a generalized statement (i.e., the lower the base-rate accuracy of any statement or group of statements, the more unique the statement

would be). Uniqueness is also used to describe the procedures whereby the student is led to *believe* that the particular generalized statements that he/she receives have resulted from the interpretation of some type of assessment process. The appropriateness and effectiveness of the procedure used by the researcher to create this illusion of uniqueness of test interpretations no doubt will affect students' ratings of generalized interpretations, as Snyder et al. indicate. However, this procedure can be and should be independent of the question whether the set of generalized interpretations that the student receives is unique or descriptive of only one person. Greene (1977) used the word *unique* in the latter sense, in that he asked students whether they realized that generalized personality statements could be applied as appropriately to one of their classmates as to themselves. Snyder et al. (1978) stated that "the *content* of the interpretive feedback *should not* be unique in order to elicit acceptance on the part of the recipient" (p. 1493). Why they should want to explicitly eliminate the only pertinent question that can be asked of generalized personality interpretations is not clear. To reiterate an earlier point, it makes little sense to ask students whether generalized interpretations are accurate. The imperative question is whether students can recognize that generalized interpretations are trivial because they can be applied to anyone, and Greene (1977) demonstrated that students can indeed recognize the triviality of generalized statements at least in one context.

The author appreciates the comments of James R. Clopton and Donald H. Baucom on a preliminary draft of this manuscript and the literary contribution of Robert H. Weiner.

Requests for reprints should be sent to Roger L. Greene, Department of Psychology, Texas Tech University, Lubbock, Texas 79409.

Snyder et al. (1978) do raise one valid methodological issue. Since a repeated-measures design was used in Greene's (1977) original study, it is necessary to conduct a similar study using either a between-subjects design or a counter-balanced repeated-measures design to determine the generality of his results. Only empirical data

can answer the methodological question that they have raised.

One final issue remains to be discussed—whether clients can provide clinicians with feedback about their personality interpretations. Snyder et al. (1978) stated that “although clinicians in actual practice may ask clients to give their acceptance reactions to feedback, they probably do not ask clients about the uniqueness of the feedback” (p. 1494). Clinicians who are interested in avoiding the pervasive influence of the Barnum effect should be specifically asking their clients the latter question. If clients can recognize whether personality interpretations are unique, which needs to be empirically tested, clients can be a valuable source of feedback to clinicians.

In concluding this reply, a quote comes to mind: “What’s gone and what’s past help/Should be past grief.” (Shakespeare, 1963, p. 65: *The Winter’s Tale*, Act III, Scene 2). Hopefully, experimenters will allow the previous approach to studying students’ acceptance of generalized personality interpretations to die a peaceful death and instead infuse some new direction into an interesting research area.

References

- Forer, B. R. The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology*, 1949, 44, 118-123.
- Greene, R. L. Student acceptance of generalized personality interpretations: A reexamination. *Journal of Consulting and Clinical Psychology*, 1977, 45, 965-966.
- Shakespeare, W. The winter's tale. In J. H. P. Pafford (Ed.), *The arden edition of the works of William Shakespeare: The winter's tale*. Cambridge, Mass.: Harvard University Press, 1963.
- Snyder, C. R. Acceptance of personality interpretations as a function of assessment procedures. *Journal of Consulting and Clinical Psychology*, 1974, 42, 150.
- Snyder, C. R., Handelsman, M. M., & Endelman, J. R. Can clients provide valuable feedback to clinicians about their personality interpretations? A reply to Greene. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1493-1495.
- Ulrich, R. E., Stachnik, T. J., & Stainton, N. R. Student acceptance of generalized personality interpretations. *Psychological Reports*, 1963, 13, 831-834.

Received February 14, 1978 ■

Premature Conclusions Regarding Black and White Suicide Attempters: A Reply to Steele

Richard C. Bedrosian and Aaron T. Beck

Department of Psychiatry
University of Pennsylvania

The conclusion reached by Steele, that his samples of black and white suicide attempters were clinically similar is questioned, as is his inference that his results require a reexamination of the supposed need for separate black and white psychologies. Specific criticisms are raised regarding some of the variables chosen by Steele for his comparisons, and the manner in which he chose to interpret his data.

As a result of his comparison of black and white suicide attempters, Steele (1977) concluded not only that the two groups were clinically quite similar but also that his data necessitated a reexamination of the contention that separate psychologies are needed for blacks and whites (Mosby, 1972). Although data contrary to the latter conclusion (which is viewed by us as premature) will no doubt be drawn from a variety of sources by other critics, the point to be made herein is that the proclamation of "no difference" between the black and white suicide attempters in Steele's sample was likewise ill-advised.

Steele concluded that the black-white differences in his sample were negligible after having observed in his comparison between the two groups that "only 11 out of the 42 clinical variables are statistically significant at the 5% level of confidence" (Steele, 1977, p. 983). Implicit in his statement is the assumption that on an *a priori* basis, many more of the variables should have revealed a difference between the two groups. In other words, Steele seems to have assumed that all or most of the variables that he examined represent equally important aspects of suicidal behavior and/or ideation, and that the way to test for differences between the two groups is to compare them on a host of variables and then assign a "box score." Apparently no effort was expended to select suicidal risk variables that reflect either observed or hypothesized differences between white and black subcultures.

What the box score approach also overlooks, of course, is that some variables may be more useful or interesting than others. With regard to suicidal behavior, "pessimism and hopelessness" might be more relevant as a predictor than "paranoid delusions" or "patient's acknowledged attempt to get some sleep," for example.

Closer examination of the indices used by Steele (1977) reveal some variables that might have been expected to bear little relation to the phenomenon of suicidal behavior. "Depressive delusions," "paranoid delusions," and "thought disorder" all presumably referred to the presence of psychotic processes on the part of the suicide attempter. Yet there is evidence that the neurotic-psychotic distinction may not be particularly relevant to the study of suicidal behavior. In a study of attempters, neurotics and psychotics showed no differences on self-report and clinical measures of suicidal intent, as well as on ratings of the lethality of their attempts (Lester & Beck, 1976). On the other hand, in view of the data supporting a relationship between alcohol intake and suicidal behavior (Mayfield & Montgomery, 1972), Steele might well have included a measure of alcohol consumption at the time of the attempt. These examples were not raised in order to second guess Steele; they have been included here in order to caution against premature closure on the issue of the black-white comparison.

If one assumes as Steele (1977) apparently does not that his data represent random results, then his findings do have some clinical relevance. Whites scored higher than blacks on ratings of depression and hopelessness, two variables that traditionally have been used to assess suicidal intent and that have consistently surfaced in the

Requests for reprints should be sent to Richard C. Bedrosian, Department of Psychiatry, Room 602, University of Pennsylvania, 133 South 36th Street, Philadelphia, Pennsylvania 19104.

literature as predictors of suicidal risk (Beck, Kovacs, & Weissman, 1975; Minkoff, Bergman, Beck & Beck, 1973). Moreover, Steele reported that blacks were less motivated to appeal for help and that they spent less time between planning the act and actually making the suicidal attempt. In trying to assess the suicidal risk in a black patient, the psychotherapist may find that the usual cues (e.g., hopelessness, depression, "cry for help") for making such judgments are less useful than if a white patient had been involved. Further, Steele's data suggest that the black patient may require a more rapid intervention in order to avoid an attempt once the suicidal ideation has been detected.

A more adequate test of whether black and white suicide attempters represent different populations should use dependent variables that are (a) known to bear a relationship to the phenomena of suicidal behavior and/or ideation and are (b) presumed to reflect some area of dissimilarity, either observed or hypothesized, between black and white subcultures. This critique raised questions about the way in which Steele (1977) satisfied the first criterion, but it seems clear that he paid no attention to the second. Some well-informed hunches about where the differ-

ences between black and white suicide attempters might lie could have provided Steele's project with a clearer focus, and perhaps, greater credibility.

References

- Beck, A. T., Kovacs, M., & Weissman, A. Hopelessness and suicidal behavior: An overview. *Archives of General Psychiatry*, 1975, 234, 1146-1149.
- Lester, D., & Beck, A. T. Suicidal behavior in neurotics and psychotics. *Psychological Reports*, 1976, 39, 549-550.
- Mayfield, D. G., & Montgomery, D. Alcoholism, alcohol intoxication, and suicide attempts. *Archives of General Psychiatry*, 1972, 27, 349-353.
- Minkoff, K., Bergman, E., Beck, A. T., & Beck, R. Hopelessness, depression and attempted suicide. *American Journal of Psychiatry*, 1973, 130, 455-459.
- Mosby, D. Toward a new specialty of black psychology. In R. L. Jones (Ed.), *Black psychology*. New York: Harper & Row, 1972.
- Steele, R. E. Clinical comparison of black and white suicide attempters. *Journal of Consulting and Clinical Psychology*, 1977, 45, 982-986.

Received December 19, 1977 ■

Steele's Reply to Bedrosian and Beck

Robert E. Steele
University of Maryland

The two basic criticisms of Bedrosian and Beck are addressed: the appropriateness of the variables chosen and the interpretation of the results with regard to black psychology. A rationale is given for the inclusion of clinical variables that were not particularly relevant to suicide attempt behavior. Two additional variables, history of alcohol and illicit drug abuse, are reported to support the similarity position between black and white suicide attempters. A final comment is made on the relevance of the need for a separate black psychology.

Bedrosian and Beck's (1978) criticism of my recent article (Steele, 1977) revolves around two issues—(a) the appropriateness of the variables chosen to examine black-white differences in suicide attempt behavior and (b) the prematurity of my claim that one empirical "box score" study is sufficient to question the viability of a black psychology. Bedrosian and Beck chose to address the first point themselves and suggested that "other critics" would attack the latter. Let me take this opportunity to counter both Bedrosian and Beck as well as to address the other critics on the second issue.

Bedrosian and Beck's (1978) first line of attack was that I assume that "all or most of the variables . . . represent equally important aspects of suicidal behavior and/or ideation" (p. 1498). I made no such assumption. In fact, the data summarized and condensed in the first paragraph of Footnote 1 (Steele, 1977, p. 984) were initially presented in three tables in the original manuscript submitted to this journal. These variables were selected after an extensive literature review of suicide attempt behavior. Key variables identified by Farberow and Schneidman (1961) and Stengel (1964) were included in this study.

The rationale for including the other "host of variables" was that in addition to focusing on suicide attempt behavior between blacks and whites, I wanted to see what an extended "clinical profile" between these two groups would look like in light of the claims made by *others* that there is a need for a separate black

psychology. This claim does not represent my point of view. Therefore, Bedrosian and Beck are incorrect to suggest that I expected "on a priori basis" black-white differences.

Bedrosian and Beck's (1978) second line of attack was my inappropriate inclusion of psychotic clinical variables (depressive delusions, paranoid delusion, and thought disorder) in light of the findings by Lester and Beck (1976). I would like to point out to my critics that my study was planned, the literature review was conducted in 1969, and the results of the Lester and Beck study were reported in 1976.

Bedrosian and Beck's third line of attack was that I committed the "sin" of omission; that is, I left out a key variable known to be related to suicide attempt behavior, namely alcohol consumption. I concede that my study would have been improved if this variable had been included. Let me hasten to add, however, that the residents did take a history of alcohol and illicit drug abuse as a part of their medical work-up. On these two variables, which were not reported in the original study, there were no statistically significant differences between black and white suicide attempters that are consistent with my interpretation that there are no meaningful "clinical" differences between these two groups.

Let me make a final comment on the second point that other critics will address. It was not my original intention to suggest that this one study is sufficient to invalidate the claim that a black psychology is needed. However, my interpretation of these results does not provide supportive evidence for a black psychology notion. I feel that my call for a reexamination of the need for a black psychology in light of empirical evidence is a valid one. A recent study

Requests for reprints should be sent to Robert E. Steele, Department of Psychology, University of Maryland, College Park, Maryland 20742.

by Steele (1978) that examined psychosocial factors in depression (a variable related to suicidal behavior) found that there were no significant relationships between race and measures of depression. Race did interact significantly with other factors such as social mobility. Upwardly mobile blacks and downwardly mobile whites showed more psychological disturbances on factors relating to depression. It is my growing conviction, which has some empirical support, that race and/or ethnic identity is a critical factor in determining behavior in complex interactions with other significant variables such as sex, religion, social class, physiological correlates, and so forth. Just because each of these factors may be important, I do not feel that it makes theoretical and/or empirical sense to elevate any of these factors to a salient status, such as a "female" and/or "black" psychology. I agree with Kluckhohn (1949) that human beings regardless of our cultural backgrounds share certain universal characteristics, that certain characteristics are shared with other members of our ethnic groups, and that we share certain characteristics as individuals. It is the goal of psychology to understand the complex interrelationships among these three factors. I

do not feel that a narrowly conceived ethnic and/or individual psychology will achieve this goal.

References

- Bedrosian, R. C., & Beck, A. T. Premature conclusions regarding black and white suicide attempters: A reply to Steele. *Journal of Clinical and Consulting Psychology*, 1978, 46, 1498-1499.
- Farberow, N. L., & Schneidman, E. S. (Eds.). *The cry for help*. New York: McGraw-Hill, 1961.
- Kluckhohn, C. *Mirror for man: The relation of anthropology to modern life*. New York: McGraw-Hill, 1949.
- Lester, D., & Beck, A. T. Suicidal behavior in neurotics and psychotics. *Psychological Reports*, 1976, 39, 549-550.
- Steele, R. E. Clinical comparison of black and white suicide attempters. *Journal of Consulting and Clinical Psychology*, 1977, 45, 982-986.
- Steele, R. E. Relationship of race, sex, social class, and social mobility to depression in normal adults. *Journal of Social Psychology*, 1978, 104, 37-47.
- Stengel, E. *Suicide and attempted suicide*. Middlesex, England: Penguin Books, 1964.

Received May 8, 1978 ■

Psychotherapy or Massage Parlor Technology? Comments on the Zeiss, Rosen, and Zeiss Treatment Procedure

Kent G. Bailey
Virginia Commonwealth University

There are numerous ethical, moral, philosophical, and social psychological issues involved in modern sex therapy. Psychologists have accorded sex therapy a warm reception into the field, but present ethical guidelines are insufficient to protect clients from psychological damage in the form of massive intrusions on privacy and reoriented moral and religious values. Further, the more explicit procedures seem to carry a message to society that "anything goes." The Zeiss, Rosen, and Zeiss procedure is used as a reference point for discussing these various issues.

In a recent review, Byrne (1977) outlined the progressive evolution of research on sexual behavior from animal studies in the early phases, to primitive cultures, abnormal sexuality, and finally up to the contemporary focus on normality emanating from Freud, Kinsey, and Masters and Johnson. Byrne stated that social psychologists were attracted to the area by virtue of revolutionary societal changes in attitudes and actual behavior, by research initiated by the Commission on Obscenity and Pornography, and by the pressing need to solve the problem of unwanted conception. Following his in-depth historical analysis of research on sexuality, Byrne concluded that social psychologists have much to gain and much to offer by providing a "warm welcome" to this fundamental area of social psychological interest.

Perusal of recent issues of the *Journal of Consulting and Clinical Psychology (JCCP)* reveals that sex research is being warmly received by clinical researchers and therapists as well, and, further, it appears that there is a parallel trend to use increasingly explicit materials, methods, and procedures in the sex laboratory. A recent *JCCP* issue, for example, included a study in which introductory psychology students viewed films of masturbation and subsequently responded to various measures of sexual arousal and affectivity (Mosher & Abramson, 1977), and another study used a behaviorally oriented masturbation procedure for treating inorgastic women

(Zeiss, Rosen, & Zeiss, 1977). Although both of these studies produced some data of interest, the explicit, high-impact treatment procedures raise numerous ethical, moral, philosophical, and potential legal issues that were not addressed, and the crucial issue of cost versus benefit was virtually ignored. In other words, there was little concern as to whether the data or treatment effects generated outweighed the potential psychological harm to the subject or client, and, more importantly, there was little concern regarding the cumulative social impact of the "anything goes" message that such procedures communicate to the subject, the client, other professionals within and out of the field of psychology, and, finally, at some point, to the public at large.

Using the Zeiss et al. (1977) study as a reference point, I would like to address some of the general issues raised above in greater detail.

The Ethical Issue

Although intangibles like "intent," "competence," and even "confidentiality" possess unclear ethical boundaries, all therapists agree on one thing—Sexual liaison between therapist and client is unethical and is to be uniformly proscribed (American Psychiatric Association, 1973; American Psychological Association, 1977). Furthermore, such contact produces negative psychological effects on the client in most instances (Taylor & Wagner, 1976) and typically represents a form of exploitation of female clients by male therapists (American Psychological Association, 1975). Indeed, sexual contact within the therapy relationship may be seen as

Requests for reprints should be sent to Kent G. Bailey, Department of Psychology, Virginia Commonwealth University, Richmond, Virginia 23284.

tantamount to breaking the incest taboo (Taylor & Wagner, 1976), for like it or not our desperate, anxious, and lonely clients cast us in a parental role, and our own needs catalyze the process.

Now, we have here an instance in which a particular form of sexual expression is condemned as morally reprehensible, contrary to client welfare, and patently unethical. It is rare that behavioral scientists are so definitive about anything, and I cannot think of any other instance in the sexual realm in which the blue pencil applies with such stringency. Given the innumerable possibilities for sexual behavior and misbehavior, it is surprising that only therapist-client sexual contact should warrant our ire. I suggest that we be fair, and, rather than "warmly welcoming" any new combination or permutation of arms, legs, and genitalia that leads to orgasm, we should use caution and discretion in intruding into the sex lives of our clientele.

Unfortunately, the Zeiss et al. (1977) study fails to exercise such caution and discretion, and what we have instead is a simplistic, functionalistic, treatment "program" that successfully yields orgasm, ignoring all the while the more profound psychological implications of the procedure. I in no way question the intent or motivation of Zeiss et al., for they behaved as ethical behavior therapists should, by using the simplest, most direct, and most effective techniques to achieve treatment goals in the shortest time possible (AABT, 1976). I do, however, question their methods and a system of professional ethics (or lack thereof) that allows therapists to declare open season on the sexually maladjusted client, with the only interdiction being that the therapist cannot join in on the fun. If sexual liaison with the client is the therapist's version of "playing house," then no-holds-barred treatment procedures like those of Zeiss et al. makes for "playing doctor," with the willing, dolllike client going along with anything—and I mean anything! The gradual fading in of Mr. A comes to mind here (Zeiss et al., 1977, p. 893), along with Mrs. A apparently thrusting and vibrating herself into a case of pneumonia, or, better still, Mrs. B's sudden, mysterious sexual longing for organic vegetables following her reading of Friday's (1974) *My Secret Garden*.

All of this would be humorous were it not so tragic. It is a mystery to me how conversational psychotherapy has made the sudden transition to massage parlor technology involving vibrators, mirrors, surrogates, and now even carrots and cucumbers! Can we say that the empirical, theoretical, and philosophical groundwork has been

laid for such procedures? Do we have data on the psychological, social, and spiritual side effects of such procedures? I think not; it is more parsimonious to assume that therapists are giving clients what they want (or think they want) whether they need it or not. Furthermore, the distinction between "legitimate" and "illegitimate" sex therapists is becoming increasingly blurred as the money pours in and the pop therapy bandwagon rolls on. The Zeiss et al. procedure represents one of the more dramatic instances of "legitimacy by acclamation," with applause supplied by the behavioral community on the one hand and the instant gratification crowd on the other. The Zeiss et al. procedure and others like it do little to allay Albee's (1977) fear that modern society is creating "an impulse-indulgent society of consumers, and psychotherapists have become the new gurus explaining life's elusive purpose" (p. 150).

Our system of professional ethics, then, should develop the breadth and depth to accommodate more than the obvious parameters of competence, confidentiality, interpersonal relations, use of tests, effectiveness of treatment, and so forth. It should also grapple with the subtler aspects of client welfare, including intrusions of privacy, indirect assaults on the client's moral and religious values, and, most importantly, abuse of the power to direct and facilitate social change. Just as Freud, Skinner, and even Spock have effected massive social change directly with new ideas and techniques, and indirectly with innumerable covert stipulations that emanate from their work, the modern sex therapist carries a powerful message to a credulous public looking to psychology for guidance. Although the field of psychology in general implicitly espouses exaggerated individualism (Hogan, 1975), liberal ideology (Ornstein, 1975), and instant gratification (Albee, 1977), it is the sex therapist who most seriously assaults the traditional institutions (see Campbell, 1975) of marriage, family, and religious values. It is time that we come to grips with these admittedly elusive side effects to therapy, and it is time that we set proper ethical guidelines and sanctions within the field before inevitable legal and professional disasters befall us from without. The recent brouhaha at the State University of New York at Albany (Smith, 1977) is a frightening illustration of the catastrophes that can befall the psychologist when ethical guidelines are ignored or are given mere token consideration.

The Legal Issue

As Tryon (1976) pointed out, concern with the ethics and legality of behavioral practices increases in proportion to the effectiveness of learning principles applied to human behavior. When the therapeutic procedures are highly specific and highly effective, the independent variables are laid bare for legal scrutiny, and this naturally leads to discussion of the whos, whens, and wheres of using such powerful procedures. The Zeiss et al. (1977) procedure ranks at the top in both specificity and "effectiveness" in the orgasmic sense, and the question arises as to how such procedures would fare in the courtroom.

Tryon (1976) reminds us that we live in an increasingly litigious society, and behaviorally oriented clinicians especially need to take necessary defensive precautions in order to avoid malpractice suits. The most important defensive precautions revolve around the subtleties of informed consent and several steps are involved: (a) The therapist lists behavioral goals requested by the client; (b) the therapist provides a statement of goals, which may or may not parallel those of the client; (c) once client and therapist goals are reconciled, the therapist makes full disclosure of treatment options, and procedures are described in detail; and finally, (d) the therapist acknowledges any reasonable *side effects* that may be expected as a result of treatment. In reading the Zeiss et al. article, I cannot tell whether any of the above precautions were taken, but I presume, as good behavioral therapists, that the authors went so far as to reconcile client and therapist goals, and hopefully the clients were presented with treatment alternatives and contingencies. However, the authors appear totally insensitive to the serious matter of side effects, and concern is instead focused on technical considerations and criteria for patient selection and rejection.

It is debatable as to whether the Zeiss et al. procedure would survive a malpractice suit by virtue of failure to fully consider potential side effects. For example, Mr. and Mrs. A were appropriately cooperative and seemed to enjoy their rather bizarre six-step program as it was in progress, but imagine what could happen a year later when Mr. A looks back and sees himself and his wife as "demeaned," "humiliated," and suffering "irreparable psychological harm" by virtue of the massive reorientation of habits, values, and self- and other-percepts that followed from treatment. Could the Zeiss et al. procedure withstand rhetorical assault from counsel for the plaintiff, who would, no doubt, dwell

on the grievous "wounding of spirit" suffered by his or her clients?

The client's right to privacy is another important issue with potential legal implications. Until sex therapy came along, the psychotherapist merely intruded into every cognitive crack and crevice possible, but now we "play doctor" to the fullest and perform behavioral "operations" that intrude into the client's private life to a degree unimaginable a decade ago. Unless the sex therapist can demonstrate that less intrusive alternate procedures are not feasible on a cost-benefit basis, then it seems to me that he or she could be subject to a malpractice suit if challenged by a client on that basis. I am unaware of any empirical research that has addressed the question of differential effectiveness vis-à-vis "hard" sex therapy procedures like those of Zeiss et al. versus "soft" conversational or less direct approaches. Unless, the hard, more intrusive approaches are found to be vastly superior in effectiveness to softer ones, then application of the former approaches would be at worst unethical and possibly illegal, and at best, just bad therapy.

The Power Issue

One of the most important and generally ignored variables in the therapy situation is that of therapist power and dominance over the client. Anyone who has supervised therapy practicum for a 2nd or 3rd-year clinical student, however, is immediately struck with the jockeying for control that goes on between this young pre-PhD and the client, and when we are informed that "rapport has been established," we know, at long last, that the neophyte therapist has finally gotten the upper hand. Were it not for the almost hypnotic power of the therapist, one wonders whether therapy would occur at all, for a great amount of what we call therapy is just plain persuasion and the selling of agendas of various sorts. It is difficult to imagine therapy proceeding on the basis of client aversion and disrespect for the therapist, and it seems to me that client cooperation is the most precious commodity in the therapy situation.

With such a gross status asymmetry and imbalanced dominance relationship, the therapist is in a position to extract levels of "obedience" that Milgram could only dream of, and such power necessarily requires judicious control and self-monitoring throughout the entire therapy process. In view of these considerations, I am not surprised that Zeiss et al. (1977) had little

difficulty in introducing the dildo in treatment, nor would I disagree that "clients will accept the procedures [the six-step program] when they are presented with an appropriate rationale" (p. 894). It might be more instructive to ask, "What will a client who needs therapy *not* do for the marketer of the product that he or she wishes to consume?" The answer—practically nothing! If anyone doubts this assertion, I suggest that they read Rosen's (1977) article dealing with the ready tendency for clients to sign away their rights to privacy under sign-away pressure, one type of which is the implicit message "do it my way or you don't get therapy." Thus, we see that mere client consent, or client capitulation as the case may be, is a suspect criterion for legitimizing a particular therapeutic intervention.

There is a related issue that I wish to bring up at this point. Despite intellectual acceptance of the trappings and regalia of the sex laboratory, I see little evidence that sex therapists wish to ply their trade on members of their own families, friends, neighbors, or loved ones. Rather, it is the impersonal stranger who is the modal consumer of the sex therapists' goods, and, as such, may be treated within the context of a dangerous underlying double standard. According to W. D. Hamilton's kinship selection theory (Barash, 1977), we tend to be more altruistic and beneficent toward those with whom we share the greatest number of genes, whereas those who share few genes with us fall in the despised, or, at best, tolerated outgroup. At the social level, we see a similar in-group versus out-group theme in Nietzsche's master morality and slave morality, a double-standard in which the aristocratic class exploits and oppresses the masses for their own good. Perhaps comparing Hamilton's theory to the therapy relationship is a false analogy, and perhaps suggesting that therapists can, in some respects, be seen as an aristocratic class ministering to the masses is a bit strong, but there is little evidence to suggest that clients are treated with the dignity and respect that we reserve for ourselves and significant others. For example, I have often seen therapists speak very cavalierly of "divorce," "masturbation," "surrogate partners," and the like in reference to clients while their own divorces and sex lives are treated with the utmost reverence. I personally believe that clients deserve the respect and reverence that we accord our loved ones, and I question whether the Zeiss et al. procedure, and similar ones, operate on that basis.

The Philosophical-Theological Issue

I believe that sex therapy is appropriate and desirable when performed by a *psychotherapist* who combines sensitivity to philosophical, moral, and theological issues with his or her armamentarium of technological skills. Without this sensitivity, poking and punching with dildos and the like becomes little more than animal husbandry at the human level. When science discards its humanity for the sake of technology, its subjects and clients become things rather than people (Goodfield, 1977), and science becomes estranged from the society that sustains it. As Goodfield so eloquently reminds us, this is nothing new, for science has, from its very beginning, been alienated from both the arts and the laity. She summarizes the historical background for the two most recurring criticisms of science: (a) First, science is cold and inhuman and does not concern itself with the needs of society; and (b) second, somehow science manages to extract the warmth and beauty of the world, and, in the process, drains itself. In concluding her essay, Goodfield asserted that science and society can no longer afford estrangement, and the scientist must work to understand the public as well as vice versa. Furthermore, the complexities of modern society disallow professional allegiance to a methodological ethic alone; rather, the scientist must work to apply "knowledge of facts in new compassionate ways" (Goodfield, 1977, p. 585).

One good way to start is by recognizing that a plurality of perspectives is needed to gain even minimal understanding of something so apparently straightforward as the masturbatory behavior of a given client. It is naive and arrogant to assume that masturbation can be instituted or facilitated in a treatment program without due consideration of the impact this may have on the moral and spiritual well-being of the client. Campbell (1975) has argued against psychotherapy's unwarranted rejection of moral tradition, and, more recently, the issue of therapeutic intrusion into the realm of religious belief has surfaced with considerable pandemonium (Cohen, 1977; McLemore & Court, 1977; Stokes, 1977). Such discussion is long overdue, and as McLemore and Court (1977) put it, "it would be professionally irresponsible of psychologists to ignore the wider implications of deciding to alter clients' belief systems" (p. 1175). Zeiss et al. may argue that none of this applies to their procedure, but knowledgeable behavior therapists (e.g., Lazarus, Mahoney, Meichenbaum) have conceded that the cognitive sphere cannot be

artificially separated from action and behavior. On the face of it, a masturbation program of any type would appear to carry innumerable hidden messages and covert stipulations—all or most of which are antithetical to Judeo-Christian traditions—and it is high time that sex therapists pay some attention to the impact that their procedures have on clients' values and religious beliefs.

Concluding Comment

As Kuhn (1970) tells us, science in its advanced stages is naturally insulated from society and social problems, and each more-or-less isolated scientific group shares common values and beliefs, and a restricted view of the world is taken for granted. It is only within such a context that the more explicit forms of sex therapy operate, hidden behind behavioristic jargon and academic degrees. Only on closer scrutiny by the disciplinary nonbeliever or the lay outsider do we see that distinctions between legitimate and illegitimate forms of sex therapy are extremely difficult to make; and only then do we see the tenuous ethical, moral, and professional nature of the enterprise.

References

- AABT Newsletter, 1976, 3, 2-4.
- Albee, G. The Protestant Ethic, sex and psychotherapy. *American Psychologist* 1977, 32, 150-161.
- American Psychiatric Association. The principles of medical ethics with annotations especially applicable to psychiatry. *American Journal of Psychiatry*, 1973, 130, 1057-1064.
- American Psychological Association. Report of the Task Force on Sex Bias and Sex-Role Stereotyping in Psychotherapeutic Practice. *American Psychologist*, 1975, 30, 1169-1175.
- American Psychological Association. Ethical standards for psychologists. *APA Monitor*, March 1977, pp. 22-23.
- Barash, D. *Sociobiology and behavior*. New York: Elsevier, 1977.
- Byrne, D. Social psychology and the study of sexual behavior. *Personality and Social Psychology Bulletin*, 1977, 3, 3-30.
- Campbell, D. T. On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist*, 1975, 30, 1103-1126.
- Cohen, R. J. Socially reinforced obsessing: A reply. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1166-1171.
- Friday, N. *My secret garden*. New York: Pocket Books, 1974.
- Goodfield, J. Humanity in science: A perspective and a plea. *Science*, 1977, 198, 580-585.
- Hogan, R. Theoretical ecocentrism and the problem of compliance. *American Psychologist*, 1975, 30, 533-542.
- Kuhn, T. S. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1970.
- McLemore, C. W., & Court, J. H. Religion and psychotherapy—Ethics, civil liberties and clinical savvy: A critique. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1172-1175.
- Mosher, D. L., & Abramson, P. R. Subjective sexual arousal to films of masturbation. *Journal of Consulting and Clinical Psychology*, 1977, 45, 796-807.
- Ornstein, A. C. The limitations of research: A viewpoint. *American Psychologist*, 1975, 30, 511-513.
- Rosen, C. E. Why clients relinquish their rights to privacy under sign-away pressures. *Professional Psychology*, 1977, 8, 17-24.
- Smith, R. J. SUNY at Albany admits research violations. *Science*, 1977, 198, 708.
- Stokes, J. B. Comment on "Socially reinforced obsessing: Etiology of a disorder in a Christian Scientist." *Journal of Consulting and Clinical Psychology*, 1977, 45, 1164-1165.
- Taylor, B. J., & Wagner, N. N. Sex between therapists and clients: A review and analysis. *Professional Psychology*, 1976, 7, 593-601.
- Tryon, W. W. Behavior modification therapy and the law. *Professional Psychology*, 1976, 7, 468-474.
- Zeiss, A. M., Rosen, G. M., & Zeiss, R. A. Orgasm during intercourse: A treatment strategy for women. *Journal of Consulting and Clinical Psychology*, 1977, 45, 891-895.

Received December 27, 1977 ■

Is Masturbation Still Wrong? Comments on Bailey's Comments

Nathaniel N. Wagner
University of Washington

Although Bailey in commenting on the Zeiss, Rosen, and Zeiss article on masturbatory techniques in sex therapy, sees ethical, moral, philosophical, social and psychological issues, he lists no specific negatives except "potential side effects." Bailey demands that therapy using directed masturbation must be "vastly superior" to conventional techniques or else it is unethical. No rationale for this discrimination is presented except potential side effects. The antiscientific nature of this argument is noted.

Bailey's (1978) comments on the Zeiss, Rosen, and Zeiss (1977) article on "Orgasm During Intercourse: A Treatment Strategy for Women" continues a long tradition of intellectual and emotional arguments against masturbation. This tradition reached its zenith in the mid-18th century when S. Tissot of France wrote "Onana, a Treatise on the Diseases Produced by Onanism" (Dearborn, 1966). Following the intellectual and moral position set forth by Tissot in 1866, a British physician, Isaac Baker Brown, described the surgical removal of the clitoris to prevent masturbation and its serious side effects. The title of his volume communicates the flavor: "On the Curability of Certain Forms of Insanity, Epilepsy, Catalepsy and Hysteria in Females" (Corea, 1977).

The essence of Bailey's (1978) criticism of Zeiss et al. is that they are insensitive to the "psychological, social, and spiritual side effects of such procedures" (p. 1503). A little later, Bailey acknowledges that "these admittedly elusive side effects to therapy" (p. 1503) have not occurred as yet. This does not prohibit Bailey from warning of the serious negatives that may accrue to "the traditional institutions of marriage, family, and religious values" (p. 1503).

Let us be clear about what Zeiss et al. (1977) recommend in their article. A six-step treatment program for women who are inorgasmic during intercourse is described. The program helps these women experience orgasm with vaginal containment of their partner's penis, whereas

previously they could only experience orgasm with manual stimulation of the clitoris. Masturbation with a dildo is used as a means of generalizing the response from clitoral stimulation to vaginal containment. There is a thoughtful discussion of the work of Masters and Johnson (1966) demonstrating the physiologic similarity of orgasm regardless of the main source of stimulation. Zeiss et al. (1977) argue that if a woman wants to experience orgasm with intercourse, this "is a justifiable behavioral goal" (p. 891). Although controversial among sex therapists, their position is a reasonable one that has many adherents. The opposing view is that all orgasms are equivalent and that it is a waste of time and may be unsuccessful to attempt to change the orgasmic pattern.

That many women seek medical and psychological assistance with orgasmic problems is well established. Masters and Johnson (1970), LoPiccolo and Lobitz (1972), Kaplan (1974), Barbach (1974), and Heiman, LoPiccolo, and LoPiccolo (1976) provide clear evidence of continued innovation, treatment, and research into this aspect of female sexual dysfunction.

Bailey's (1978) critique suggests a return to more conventional talking therapy as less dangerous of potentially serious side effects. He talks about the harm that can befall therapists if they do not use "caution and discretion in intruding into the sex lives of our clientele" (p. 1503). It is hard to understand how one can intrude into the sex life of someone who comes for assistance with a sexual problem.

In an interesting turn for one who argues from an ethical and moral stance, Bailey (1978) criticizes sex therapists for not using these techniques on "their own families, friends, neighbors, or loved ones" (p. 1505). Apparently, the

Nathaniel N. Wagner died on June 13, 1978.
Requests for reprints should be sent to Martha A. Perry, Department of Psychology, University of Washington, Seattle, Washington 98195.

need for the therapist to be as objective as possible eludes him. It is an elementary fact of therapeutic practice, as medicine has long acknowledged, that it is foolish and dangerous to be professionally involved with persons with whom one has a preexisting emotional bond.

The Story of Onan

When the arguments that Bailey raises are carefully examined, it is clear that masturbation, and its presumed serious side effects, is the central issue. It may be helpful to examine the historical origin of our attitudes toward masturbation. Within the western tradition, the Old Testament Biblical story of Onan (Genesis, 38, 6) is central. Onan refused to impregnate his older brother's widow, practicing coitus interruptus instead, which brought the wrath of God and Onan's death. The biblical scholars' argument as to whether Onan's sin was in refusing to obey the levirate requirement to give seed to his brother or whether it was in the spilling of his seed on the ground is not at issue here (Bullough, 1976). The word *onanism* has historically been synonymous with masturbation (Freud, 1927), and it is currently the word for masturbation in German and other European languages. Clearly, Onan's spilling of the seed on the ground soon evolved to other forms of spilling seed on the ground, that is, masturbation. The masculine bias should be noted.

Within the confines of a propopulation stance, which was rational and justifiable in Biblical times, masturbation threatened the survival of the species. High infant mortality and an overall high death rate helped to develop a pronatalist policy of severely censuring any nonreproductive sexual behavior (i.e., masturbation, homosexuality, and oral sex). This is the context in which Bailey (1978) seeks to "protect clients from psychological damage in the form of massive intrusions on privacy and reoriented moral and religious values" (p. 1505).

It is interesting that on January 22, 1976, Pope Paul VI reasserted the unequivocal opposition of the Roman Catholic Church to masturbation: "The moral sense of the faithful have declared without hesitation that masturbation is an intrinsically and seriously disordered act" (Kosnik, 1977, p. 306). Religious thought, however, is not uniform in its condemnation of masturbation and other nonreproductive acts. There is a growing sense of the need to reexamine traditional views (The United Church of Christ, 1977).

From a point of view that sees overpopulation as the Number 1 problem of the world

(Hardin, 1968), one could argue that masturbation is a constructive way of dealing with sexual interest. To continue to perceive masturbation as intrinsically psychologically damaging is archaic and simply not true. If there are serious side effects, where is the scientific literature documenting these effects? Bailey has not made the case.

The Ethical Issue

Having responded in a general way to the highly emotionally charged criticisms of Bailey, I would like to respond specifically to each issue raised. The first is the ethical one. Bailey (1978) states that Zeiss et al. are allowed "to declare open season on the sexually maladjusted client, with the only interdiction being that the therapist cannot join in on the fun" (p. 1503). The fact that all therapies include value judgments is not new and has been thoughtfully discussed by London (1964). Behavioral therapists are probably less guilty of making value judgments in the guise of therapy than such nonbehavioral therapists as psychoanalysts or Gestalt or humanistic therapists. A thoughtful discussion of the ethical problems in behavior therapy is provided by Goldfried and Davison (1976). Zeiss et al. are not any more unethical in using masturbation as a therapeutic technique than scores of talking therapists who have used their client's masturbatory fantasies in therapy. Zeiss et al. are probably even more effective in helping the client reach the goal that brought the person into treatment.

The Legal Issue

Bailey's (1978) argument here is that the techniques of Zeiss et al. would probably not "survive a malpractice suit by virtue of failure to fully consider potential side effects" (p. 1504). He suggests that clients might look back a year later and find themselves "demeaned," "humiliated," and "suffering irreparable psychological harm." Here we have rhetoric about potential side effects without data. There is no empirical substantiation other than the imagined plea of a counsel concerning "the grievous wounding of spirit." Every one of these legal issues was raised against contraceptive practice, and if persons like Margaret Sanger (Reed, 1978) had not been willing to risk the wrath of others, we would still be in the Dark Ages on that issue. Despite the proliferation of sex therapists, the only legal issues raised in the courts have been those of the use of surrogates and the clearly

unethical practice of sexual relations between therapists and clients (Taylor & Wagner, 1976). Bailey's fears of legal problems are just that—fears.

The privacy issue raised in Bailey's (1978) discussion of legal issues is hard to understand. He argues that unless the "hard" behavioral techniques are "found to be vastly superior" to the "soft" conversational ones, the application of these techniques would be "at worst unethical and possibly illegal, and at best, just bad therapy" (p. 1504). No reason is set forth as to why these techniques must be "vastly superior" rather than just equal. This is a strange double standard—that techniques that use directed masturbation must be vastly superior or else they are unethical. Bailey does not offer a single reason for the necessity of the different standard except the vague argument concerning potential side effects.

The Power Issue

There is not much need to comment here, for Bailey has simply stated that therapists have more power than clients, which is certainly true. He suggests that Zeiss et al. misuse their power, again without documentation except the previously noted assertion that they do not practice on their friends, lovers, and spouses.

The Philosophical-Theological Issue

The sum of Bailey's (1978) argument is that "on the face of it, a masturbation program of any type would appear to carry innumerable hidden messages and covert stipulations—all or most of which are antithetical to Judeo-Christian traditions" (p. 1506). There it is! "On the face of it," masturbation is wrong. Anyone who recommends it, even if it helps people to become well-functioning sexual human beings is guilty of defective professional judgment or to use Bailey's term is *unethical*.

Concluding Comment

Bailey has raised a number of historic concerns about nonreproductive sexual behavior in expressing concern about the use of masturbatory techniques in therapy. His argument is entirely based on potential dangers and on highly charged value judgments. His position is antiscientific, in that it recommends that no new technique be considered until it is proved to be "vastly superior" to conventional techniques. Interestingly, Bailey does not ask for the needed research to be conducted; he argues for sup-

pression of procedures that he finds unacceptable.

References

- Bailey, K. G. Psychotherapy or massage parlor technology? Comments on the Zeiss, Rosen, and Zeiss treatment. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1502-1506.
- Barbach, L. G. Group treatment of preorgasmic women. *Journal of Sex and Marital Therapy*, 1974, 1, 134-145.
- Bullough, V. L. *Sexual variance in society and history*. New York: Wiley, 1976.
- Corea, G. *The hidden malpractice: How American medicine treats women as patients and professionals*. New York: Morrow, 1977.
- Dearborn, L. W. Masturbation. In M. F. DeMartino (Ed.), *Sexual behavior and personality characteristics*. New York: Grove, 1966.
- Freud, S. Some psychological consequences of the anatomical distinction between the sexes. *International Journal of Psychoanalysis*, 1927, 8, 133-142.
- Goldfried, M. R., & Davison, G. C. *Clinical behavior therapy*. New York: Holt, Rinehart & Winston, 1976.
- Hardin, G. The tragedy of the commons. *Science*, 1968, 162, 1243-1248.
- Heiman, J., LoPiccolo, L., & LoPiccolo J. *Becoming orgasmic: A sexual growth program for women*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.
- Kaplan, H. S. *The new sex therapy*. New York: Brunner/Mazel, 1974.
- Kosnik, A. *Human sexuality: New directions in American Catholic thought*. New York: Paulist Press, 1977.
- London, P. *The modes and morals of psychotherapy*. New York: Holt, Rinehart & Winston, 1964.
- LoPiccolo, J., & Lobitz, W. C. The role of masturbation in the treatment of orgasmic dysfunction. *Archives of Sexual Behavior*, 1972, 2, 163-171.
- Masters, W. H., & Johnson, V. E. *Human sexual response*. Boston: Little, Brown, 1966.
- Masters, W. H., & Johnson, V. E. *Human sexual inadequacy*. Boston: Little, Brown, 1970.
- Reed, J. *The birth control movement and American society since 1830*. New York: Basic Books, 1978.
- Taylor, B. J., & Wagner, N. N. *Sex between therapists and clients: A review and analysis*. *Professional Psychology*, 1976, 7, 468-474.
- The United Church of Christ. *Human sexuality: A preliminary study*. New York: United Church Press, 1977.
- Zeiss, A. M., Rosen, G. M., & Zeiss, R. A. Orgasm during intercourse: A treatment strategy for women. *Journal of Consulting and Clinical Psychology*, 1977, 45, 891-895.

Received March 2, 1978 ■

Ethical and Professional Issues in Sex Therapy: Comments on Bailey's "Psychotherapy or Massage Parlor Technology?"

G. Terence Wilson
Rutgers—The State University

Sex therapy should be conducted by a skilled therapist who is sensitive to the professional and ethical issues that are inherent in treatment. Potentially adverse side effects of therapy should be continually assessed. To date, however, the findings from sex therapy, including the Zeiss, Rosen, and Zeiss procedure, show predominantly positive consequences. It is imperative that the fully informed client have decision-making primacy in setting treatment goals. Value-free therapy does not exist, and therapists must be careful not to impose their own personal biases. The principle of the least intrusive treatment alternative and the nature of intrusiveness are discussed.

In his response to Zeiss, Rosen, and Zeiss (1977), Bailey (1978) raises some important ethical and professional issues that relate to psychotherapy in general and sex therapy in particular. It is therefore particularly unfortunate that Bailey does more to distort than to clarify the critical issues in a reply that contains irrelevant, arbitrary, and extremist position statements that are empirically unsupported. Nor is a dispassionate appraisal of the issues helped by Bailey's frequently sarcastic tone, which is uncalled for. The following is a brief consideration of some of the points raised by Bailey's objections to Zeiss et al.'s procedures.

Therapeutic Ethics

It is fundamental to the behavioral treatment of sexual (or any other) problems that the client should have the major say in setting the goals of therapy. Accordingly, it is imperative that the fully informed client consents to and participates in goal setting. Informed consent is defined by the criteria of knowledge, voluntariness, and competency (Friedman, 1975). The clients described by Zeiss et al. (1977) sought therapy on a voluntary basis and certainly seem to have been competent to participate in the decision-making process. Along with Bailey, I

assume that they were given a full description and explanation of the treatment procedures. No less is required in order for such methods to be carried out successfully. However, Bailey (1978) criticizes Zeiss et al. for being "totally insensitive to the serious matter of side effects" (p. 1503).

Side Effects of Therapy

Accountability in therapy requires adequate assessment of the behavior that is the target of treatment. Side effects, or changes in behaviors that are not directly targeted for change, are possible in all forms of psychological treatment, including behavior therapy. Accordingly, relevant, related behaviors should be assessed in addition to the specifically treated behavior. Some of these changes may be judged to be desirable, whereas others may not. In behavior therapy, more often than not, these side effects have been positive (Kazdin & Wilson, 1978b). A problem faced by all practitioners is how widely to cast this assessment net in monitoring concomitant or generalized behavior change. Of course, one can always resort to the therapist's clinical acumen and sensitivity in detecting untoward (or positive) consequences of intervention. Although the therapist's judgment is not to be overlooked, more systematic measurement of the broad effects of treatment outcome is needed.

Although no formal guidelines exist, two general classes of behaviors would appear to be directly related to the specific goal of facilitat-

Requests for reprints should be sent to G. Terence Wilson, Graduate School of Applied and Professional Psychology, Rutgers—The State University, Busch Campus, Box 819, Piscataway, New Jersey 08854.

ing coital orgasm in the cases in point. One is the frequency and nature of the clients' sexual behavior as a whole. The other concerns the quality of the clients' emotional and interpersonal relationships with their partners. In this connection it should be noted that Zeiss et al. (1977) obtained pretreatment and posttreatment measures of both marital adjustment and mutual sexual behaviors using widely accepted measurement inventories (Locke & Wallace, 1959; LoPiccolo & Steger, 1974). The increases in marital satisfaction reported by Zeiss et al. are consistent with evidence from other sex therapy programs that have included directed masturbation training (Leiblum, Rosen, & Pierce, 1976). Since *both* partners were actively involved in the treatment program, the therapist had access to another source of clinical data pertaining to treatment effects and was in a favorable position to detect any adverse impact on the clients' interpersonal relationships. To the degree that the therapist, supported by evidence from the inventories, was unable to notice any negative side effects, Bailey's objection is undermined.

One of the problems with Bailey's (1978) reply is that it is consistently negative and does not offer any concrete constructive leads as to how to conduct ethically responsible therapy. For example, he fails to specify *what* side effects are to be feared. Repeatedly, he issues vague, value-laden caveats about sex therapists assaulting "the traditional institutions of marriage, family, and religious values" (p. 1503). He ignores the evidence showing that exposure to pornography does not appear to alter sexual morality (Commission on Obscenity and Pornography, 1970), that the direct behavioral treatment of sexual disorders enhances marital satisfaction and often prevents otherwise certain divorce and family disruption (e.g., Masters & Johnson, 1970), and that serious negative side effects of such therapy have yet to be reported. Quoting questionably relevant position statements by Albee (1977) and by Campbell (1975) is not good enough. At worst it amounts to ill-concealed moralizing.

Nor does Bailey (1978) specify *how* potential side effects are to be assessed or under what conditions sex therapy is ever to be appropriate. Rather, he appeals to therapy performed by "a *psychotherapist* who combines sensitivity to philosophical, moral, and theological issues with his armamentarium of technological skills" (italics added, p. 1505). Who would disagree with this? Reflecting the view of behavior therapists engaged in the treatment of sexual dysfunction, I have previously emphasized that effective

treatment

involves more than instruction in the art of body massage and the use of vibrators. Inadequate interpersonal relationships and lack of communication are more often than not the reasons for sexual distress. Accordingly, sex therapy requires trained therapists who are skilled and experienced. (Franks & Wilson, 1977, p. 401)

Zeiss et al. (1977) explicitly caution that therapists should remain sensitive to these broader and more subtle issues of marital (interpersonal) interaction and self-esteem. Sex therapy does not obviate the need to consider these factors carefully. It often is directly responsible for improving marital harmony, enhancing self-esteem, and providing greater self-fulfillment (Annon, 1974; O'Leary & Wilson, 1975).

If this is so, what then does Bailey find so objectionable? It appears that it is both the goal and the method of Zeiss et al.'s (1977) therapy.

Who Sets the Goals of Therapy?

Behavior therapy requires that the *client* have decision-making primacy in setting treatment goals. The therapist's role is to assist clients in evaluating the probable consequences of different courses of action (Bandura, 1969). In this process the therapist inevitably influences the client (Wilson & Evans, 1976). In doing so, it is crucial that the therapist's biases be recognized and honestly declared. Particular care should be taken in helping the client to differentiate between advice and information that has some empirical basis (in sex therapy, e.g., see Annon, 1974). This candor, in addition to the description of *explicit* treatment methods directed toward *specific* goals in a manner that allows *continual assessment* of progress (accountability), renders viable the notion of informed consent in clients such as those described by Zeiss et al. (1977). In this context the client's goal is to be respected. If the therapist is unwilling to cooperate with the client for either personal or professional reasons, he/she should say so and refer the client elsewhere.

This procedure, with which Zeiss et al.'s (1977) therapy seems to be consistent, can be contrasted with Bailey's caricature and implicit value judgments. His view is that "open season" is being declared on the client; that the treatment is one in which "anything goes" and "no holds are barred"; that therapists like Zeiss et al. are "playing doctor" with "dolllike" clients. Readers may judge for themselves how accurate a representation this is of the Zeiss

et al. procedure in particular or in sex therapy in general. It is my own view that these hyperbolic horrors tell us more about Bailey's personal values than about the treatment and its perception by, and effects on, the clients. Of course, Bailey, or anyone else for that matter, is entitled to be upset by the goals of sex therapy (in this case, the mutual enjoyment of masturbation and coital orgasms by a consenting husband and wife that is made possible by a direct behavioral method). Others, including sex therapists, will find this unobjectionable—even desirable. What is important is that neither view be imposed on the client.

There is a danger that in his no doubt well-intentioned desire to protect clients, Bailey is encouraging the subtle imposition of a particular set of values on clients. The theme that "the therapist knows best" and should be careful before condoning the self-indulgence in his or her clients' runs throughout Bailey's (1978) reply. Nowhere is it more obvious than when he objects to "therapists . . . giving clients what they want (*or think they want*) whether they need it or not" (emphasis added, p. 1503). The implication is clear: The therapist becomes the arbiter of what the client "really" wants and what he/she really needs. Among other contemporary treatment approaches, behavior therapy, in principle, involves a deliberate attempt to avoid this insidious and patronizing attitude. Therapists would do well to take seriously Michel's (1977) reminder that our clients "are the best experts on themselves and are eminently qualified to participate in the development of descriptions and predictions—not to mention decisions—about themselves" (p. 249).

Selecting Treatment Methods: The Criteria of Effectiveness and Intrusiveness

Effectiveness. Given effective methods, ethical and legal considerations dictate that the least intrusive/restrictive alternative be used (cf. Friedman, 1975). In making this point, Bailey concludes that there is no empirical evidence demonstrating the superiority of what he calls "hard" sex therapy procedures over "soft," indirect methods. (For a comprehensive and critical analysis of the evidence bearing on the efficacy of the direct treatment of sexual dysfunction, the reader is referred to Franks & Wilson, 1977; Kazdin & Wilson, 1978b; and Marks, 1976.) Suffice it to state here that there are studies indicating the greater efficacy of direct (behavioral) compared to indirect (verbal psychotherapy) techniques (Lazarus, 1961;

Marks, 1976; Obler, 1973) or a waiting-list control group (Munjack et al., 1976). Moreover, the more direct ("harder") the method, the greater the apparent efficacy (Kockott, Dittmar, & Nusselt, 1975; Mathews et al., 1976). This latter tendency is consistent with the more general finding that performance-based (direct) methods are more effective than those that rely on verbal or vicarious procedures (cf. Bandura, 1977a; Wilson, 1978).

There are as yet no unambiguous data showing that orgasmic reconditioning significantly alters sexual behavior (Conrad & Wincze, 1976). However, uncontrolled clinical reports continue to suggest that it may be a useful method (e.g., Lobitz & LoPiccolo, 1972; Wilson, 1973). Although most of the evidence for the efficacy of behavioral methods rests primarily on uncontrolled clinical reports, evaluation of the rules of evidence must take into account the fact that the consistently high success rates that have been reported by widely differing programs are unprecedented (e.g., Kaplan, 1974; Masters & Johnson, 1970). Nothing like it has ever been reported even with highly selected clients in uncontrolled clinical reports (O'Leary & Wilson, 1975). Although the final verdict must await the appropriate controlled outcome studies, the available evidence unquestionably indicates that direct performance-based methods are the *only* reasonable alternative for treating most forms of sexual dysfunction.

Intrusiveness. Bailey's assumption is that direct sex therapy methods such as orgasmic reconditioning are highly intrusive while traditional psychotherapy is not. The issue is more complex than this simplistic dichotomy implies, however. Orgasmic reconditioning is a well-specified technique that is *self-administered* by two informed sexual partners. It has a limited goal (orgasm); its effects on sexual behavior are reasonably predictable and clearly observable to the client. Is this what intrusiveness means? Simply because it is a direct behavior change method that involves sexuality does not necessarily make a technique intrusive.

Compare these procedural criteria with traditional psychotherapy. The success of psychodynamic therapy is predicated on the development of a workable transference relationship.¹ During

¹ Interestingly, a major reason why Masters and Johnson (1970) emphasized the importance of a dual-sex therapy team was to avoid the development of a transference relationship. Their treatment is geared to enhancing emotional communi-

the course of this intense emotional relationship, unconscious thoughts, forbidden impulses, hidden fantasies, and a wealth of deeply intimate material are probed. Therapy has the relatively vague goal of insight without specific operational referents that are immediately observable to the client (cf. Bandura, 1969). Could this not be construed as intrusive? Is this approach not more likely to delve into "life's elusive purpose" than the limited attempt to enhance orgasm during coitus? The answers to these questions will vary depending on the specific circumstances of each case and cannot be brushed aside by oversimplifications about what at face value seems to be the case.

Causal Models and the Therapeutic Relationship

Bailey (1978) portrays therapy as a process in which extremely powerful treatment methods are administered to dolllike clients by omnipotent therapists who "extract levels of 'obedience' that Milgram could only dream of" (p. 1504). Although this view is occasionally shared by some proponents and opponents of behavior modification alike, it fares poorly under critical scrutiny. Bandura (1978) has discussed the conceptual inadequacies of this sort of unidirectional causal model of human behavior, making a compelling case for the reciprocal determinism of behavioral influence. The therapist's influence may be considerable, but it is far from total. (See Davison, 1973, and Wilson & Evans, 1976, for a fuller discussion of the ways in which the therapist may influence the client, how unilateral therapist control is limited, and why self-regulated change is preferable to external influence.) It is more realistic, not to mention more humble, to note that the therapist is more a consultant than a controller, skillfully directing consciously involved clients in active, self-regulated problem-solving strategies instead of dominating puppetlike figures through "almost hypnotic power."

Conjuring up images of automatic conditioning, obedience training, and master-slave relationships makes for lively polemics but faulty psychology (Bandura, 1977b). The sense of this discussion can be summarized by observing that considerations of both ethics and efficacy require a therapeutic relationship in which the client is an active, self-directed participant.

cation between the clients themselves. This would be impeded by the client forming an emotional attachment to the therapist.

Concluding Comments

Bailey's response contains several irrelevant and arbitrary assertions that despite his disclaimer, could be read to impugn the professional ethics of sex therapists. Two of the more obvious examples can be mentioned. First, it is difficult to see how the issue of sexual contact between therapist and client is relevant to a serious analysis of the Zeiss et al. (1977) procedure. This is a mischievous juxtaposition that may mislead by "guilt through contiguity." Second, there is the allegation that sex therapists hold to a double standard of ethics in practicing massage parlor technology on clients who are strangers while shunning the same methods with their "loved ones." In fact, I know of cases that contradict this claim. If the reader finds my own subjective observation less than convincing, then my point about Bailey's speculation has been made.

Unquestionably, sex therapists should monitor both the direct and indirect effects of their treatment methods on clients' functioning. This attention to potential side effects of therapy should be part of an expanded evaluation of psychotherapy within a broader set of criteria than has usually been the case up to now (cf. Kazdin & Wilson, 1978a). The consequences of a pluralistic society are that different cultural groups will have different social values that are sometimes difficult to reconcile. There is no such thing as value-free therapy, and the impact of the therapeutic process on personal and social mores requires constant attention and analysis. In part, the resolution of these thorny issues will depend on their public airing as in the present interchange of ideas.² However, it is my contention that the judicious use of direct behavioral methods for the treatment of sexual dysfunction by a skilled therapist who is sensitive to these ethical issues constitutes sound clinical practice. The benefits to our clients are often considerable, and I suggest that the onus is now on those who would oppose these methods to show sufficient cause for continued recalcitrance.

² The questions of ethics, social norms, and the goals of treatment in sex therapy have been considered elsewhere (cf. Davison & Wilson, 1974; Garfield, 1974; Kohlenberg, 1974; Strupp, 1974). The present discussion should be viewed in this broader context of ethics and sexual behavior change.

References

- Albee, G. The Protestant Ethic, sex, and psychotherapy. *American Psychologist*, 1977, 32, 150-161.
- Annon, J. S. *The behavioral treatment of sexual problems* (Vol. 1). Honolulu: Kapiolam Health Sciences, 1974.
- Bailey, K. G. Psychotherapy or massage parlor technology? Comments on the Zeiss, Rosen, and Zeiss treatment procedure. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1502-1506.
- Bandura, A. *Principles of behavior modification*. New York: Holt, Rinehart & Winston, 1969.
- Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 1977, 84, 191-215. (a)
- Bandura, A. *Social learning theory*. Englewood Cliffs, N.J.: 1977. (b)
- Bandura, A. The self system in reciprocal determinism. *American Psychologist*, 1978, 33, 344-358.
- Campbell, D. T. On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist*, 1975, 30, 1103-1126.
- Commission on Obscenity and Pornography. *The report of the commission on obscenity and pornography*. New York: Bantam Books, 1970.
- Conrad, S. R., & Wincze, J. P. Orgasmic reconditioning: A controlled study of its effects upon the sexual arousal and behavior of adult male homosexuals. *Behavior Therapy*, 1976, 7, 155-166.
- Davison, G. C. Counter-control in behavior modification. In L. A. Hammerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, practice*. Champaign, Ill.: Research Press, 1973.
- Davison, G. C., & Wilson, G. T. Goals and strategies in behavioral treatment of homosexual pedophilia: Comments on a case study. *Journal of Abnormal Psychology*, 1974, 83, 196-198.
- Franks, C. M., & Wilson, G. T. *Annual review of behavior therapy: Theory and practice* (Vol. 5). New York: Brunner/Mazel, 1977.
- Friedman, P. R. Legal regulation of applied behavior analysis. *Arizona Law Review*, 1975, 17, 39-104.
- Garfield, S. L. Values: An issue in psychotherapy: Comments on a case study. *Journal of Abnormal Psychology*, 1974, 83, 202-203.
- Kaplan, H. S. *The new sex therapy*. New York: Brunner/Mazel, 1974.
- Kazdin, A. E., & Wilson, G. T. Criteria for evaluating psychotherapy. *Archives of General Psychiatry*, 1978, 35, 407-418. (a)
- Kazdin, A. E., & Wilson, G. T. *Evaluation of behavior therapy: Issues, evidence, and research strategies*. Cambridge, Mass.: Ballinger, 1978. (b)
- Kockott, G., Dittmar, F., & Nusselt, L. Systematic desensitization of erectile impotence: A controlled study. *Archives of Sexual Behavior*, 1975, 4, 493-500.
- Kohlenberg, R. J. Treatment of a homosexual pedophilic using in vivo desensitization: A case study. *Journal of Abnormal Psychology*, 1974, 83, 192-195.
- Lazarus, A. A. Group therapy of phobic disorders by systematic desensitization. *Journal of Abnormal and Social Psychology*, 1961, 63, 504-510.
- Leiblum, S. R., Rosen, R. C., & Pierce, D. Group treatment format: Mixed sexual dysfunctions. *Archives of Sexual Behavior*, 1976, 5, 313-322.
- Lobitz, W. C., & LoPiccolo, J. New methods in the behavioral treatment of sexual dysfunction. *Journal of Behavior Therapy and Experimental Psychiatry*, 1972, 3, 265-271.
- Locke, H. J., & Wallace, K. M. Short marital adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living*, 1959, 21, 251-255.
- LoPiccolo, J., & Steger, J. C. The Sexual Interaction Inventory: A new instrument for assessment of sexual dysfunction. *Archives of Sexual Behavior*, 1974, 3, 585-595.
- Marks, I. M. Management of sexual disorders. In H. Leitenberg (Ed.), *Handbook of behavior modification and behavior therapy*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.
- Masters, W., & Johnson, V. *Human sexual inadequacy*. Boston: Little, Brown, 1970.
- Mathews, A. M., et al. The behavioural treatment of sexual inadequacy: A comparative study. *Behaviour Research and Therapy*, 1976, 14, 427-436.
- Mischel, W. On the future of personality assessment. *American Psychologist*, 1977, 32, 246-254.
- Munjack, D., et al. Behavioral treatment of orgasmic dysfunction: A controlled study. *British Journal of Psychiatry*, 1976, 129, 497-502.
- Oblen, M. Systematic desensitization in sexual disorders. *Journal of Behavior Therapy and Experimental Psychology*, 1973, 4, 93-101.
- O'Leary, K. D., & Wilson, G. T. *Behavior therapy: Application and outcome*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.
- Strupp, H. H. Some observations on the fallacy of value-free psychotherapy and the empty organism: Comments on a case study. *Journal of Abnormal Psychology*, 1974, 83, 199-201.
- Wilson, G. T. Innovations in the modification of phobic behaviors in two clinical cases. *Behavior Therapy*, 1973, 4, 426-430.
- Wilson, G. T. Cognitive behavior therapy: Paradigm shift or passing phase? In J. P. Foreyt & D. Rathen (Eds.), *Cognitive behavior therapy: Research and application*. New York: Plenum Press, 1978.
- Wilson, G. T., & Evans, I. M. Adult behavior therapy and the therapist-client relationship. In C. M. Franks & G. T. Wilson (Eds.), *Annual review of behavior therapy: Theory and practice* (Vol. 4). New York: Brunner/Mazel, 1976.
- Zeiss, A. M., Rosen, G. M., & Zeiss, R. A. Orgasm during intercourse: A treatment strategy for women. *Journal of Consulting and Clinical Psychology*, 1977, 45, 891-895.

Comments on Levin et al. and Rosen and Kopel: Internal and External Validity Issues

Gary M. Farkas
University of Hawaii

This comment discusses problems of internal and external validity relevant to applied laboratory research in the modification of sexual behaviors. Used as examples are two recent case studies published in this journal. Suggestions are presented for improving the internal validity and the generalizability of studies using the laboratory paradigm.

In a recent issue of this journal, Levin, Barry, Gambaro, Wolfensohn, and Smith (1977) and Rosen and Kopel (1977) presented attempts to modify sexual response in two patients. Levin et al. sought to reduce the degree of penile response to pedophilic stimuli in a man with a 19-year history of pedophilic behavior. Rosen and Kopel attempted to modify tumescence in a man who had exhibited both transvestite and exhibitionistic behavior for 30 years. Each group of authors reported apparently successful treatments with regard to those measures assessed.

The issue for discussion in this comment will be the selection of dependent variables, both during baseline measurement and with regard to the wider issues of external validity. Although the previously cited articles are the subjects of this reply, they exemplify other studies that have similar shortcomings.

The primary purpose of baseline measurement is to provide a reference by which treatment efficacy can be evaluated. Thus, pretreatment measures provide an internal validation of an intervention's effect. One might question, however, the relevance of the pretreatment conditions used in these studies as a check for internal validity. In this discussion, I will examine only one of the measures used—percentage of full erection to deviant stimuli.

In Levin et al. (1977) and Rosen and Kopel (1977) the primary threat to internal validity was the ability of males to possess inhibitory

control over tumescence. This skill has been shown by normal subjects (Henson & Rubin, 1971), and may result from cognitive mediation (Farkas, Sine, & Evans, in press). Both Levin et al. and Rosen and Kopel made a crucial error in assuming that incarcerated patients subject to contingencies of the judicial system would provide truthful data. Since this is unlikely, some measure of deception is needed. For these particular patients, the index would be the degree to which they possessed voluntary control of tumescence to problematic stimuli.

If, for example, a patient exhibits the ability to suppress arousal to 50% of full tumescence, should not therapy be designed to enhance this skill and assessment be formulated to measure the effect of therapy above and beyond that ability presented de novo? To my knowledge, no single case or group study has taken this voluntary control issue into account.

There are additional problems of external validity that necessitate response. As Rosen and Kopel (1977) noted, strain gauge measurement is a precise and reliable indicant of physiological sexual arousal; however, the degree to which this measure provides externally valid evidence of behavior change is subject to several challenges. First, whereas tumescence may provide the first and most reliable physiological index of arousal (Zuckerman, 1971), it does not provide a measure congruent with self-report. Farkas et al. (in press) have demonstrated that under ideal conditions, the correlation between these measures averages .46, with considerable variability. Moreover, these authors maintain that a high degree of subjective arousal may be reported while minimal tumescence is evident.

Second, to what degree is penile arousal an essential criterion of sexual behavior? Stated differently, is there evidence that modification

The author is indebted to Raymond C. Rosen and Jack S. Annon for their contributions to his learning experience.

Requests for reprints should be sent to Gary M. Farkas, Department of Psychology, University of Hawaii, 2430 Campus Road, Honolulu, Hawaii 96822.

of penile responsivity in laboratory settings predicts overt behavioral changes in extralaboratory environments? Much data, both in biofeedback and aversion therapy studies of sexual problems, indicate that this is not necessarily the case (e.g., Barlow, Agras, Abel, Blanchard, & Young, 1975).

A final issue concerning generality regards the external validity of the treatment stimuli compared to stimuli encountered in the natural setting. Certainly, slides of nude girls and videotapes of transvestism-exhibitionism are unlike those stimuli evoking problematic behaviors in the extralaboratory situation. Thus, just as we have witnessed a generalization gradient to non-treatment slides (e.g., Levin et al., 1977), how steep will this gradient become when extralaboratory cues are assessed?

The present experimental literature suggests extreme caution in interpreting studies that examine reduced tumescence response to problematic stimuli as a function of some treatment. However, if one continues to accept the laboratory paradigm as having utility in treatment studies, what modifications in procedure might prove useful?

First, given the finding of unreliability between self-report and tumescence, an assessment of this factor for each patient would be valuable during the prebaseline phase to establish an upper limit of confidence in the generalizability of the data.

Second, therapists should assess the degree of inhibitory control that clients possess at prebaseline measurement. For the incarcerated patient with a high degree of voluntary control, the internal validity of any treatment effect measured by this method would be seriously questioned.

Third, the nature of the arousing stimuli used should be modified. Most often, arrests are made not because of arousal to media presentations in laboratories but because of problematic behaviors elicited by highly individual cues in the external setting. At this juncture, Abel's method of cue generation (Abel, Blanchard, Barlow, & Mavissakalian, 1975) appears to hold the most promise, at least when tumescence is the focus of intervention.

Finally, tumescence responses rarely appear to constitute the only behavior leading to in-

carceration in cases of sexual problems. Rather, a complex series of motor behaviors (e.g., fondling young children) appears to contribute to, or wholly constitute the basis for, society's unwillingness to accept certain aberrations. Thus, should not these motor behaviors be appropriate targets for direct intervention? Targeting penile response without concomitant attempts to modify cognitive and behavioral factors that may co-occur or predominate as the basis for maladaptive sexual behaviors does not appear to hold promise for maintenance and generalization of appropriate responses in the natural situation.

Although direct measurement of sexual arousal has proved a welcome innovation, the limitations of the method must be recognized. With continued discussion and experimental refinements, it is hoped that the validity of this method will improve and that more effective treatments will be developed.

References

- Abel, G. G., Blanchard, E. B., Barlow, D. H., & Mavissakalian, M. Identifying specific erotic cues in sexual deviations by audiotaped descriptions. *Journal of Applied Behavior Analysis*, 1975, 8, 247-260.
- Barlow, D. H., Agras, W. S., Abel, G. G., Blanchard, E. B., & Young, L. D. Biofeedback and reinforcement to increase heterosexual arousal in homosexuals. *Behaviour Research and Therapy*, 1975, 13, 45-50.
- Farkas, G. M., Sine, L. F., & Evans, I. M. The effects of distraction, performance demand, stimulus explicitness and personality on objective and subjective measures of male sexual arousal. *Behaviour Research and Therapy*, in press.
- Henson, D. E., & Rubin, H. B. Voluntary control of eroticism. *Journal of Applied Behavior Analysis*, 1971, 4, 37-47.
- Levin, S. M., Barry, S. M., Gambaro, S., Wolfensohn, L., & Smith, A. Variations of covert sensitization in the treatment of pedophilic behavior: A case study. *Journal of Consulting and Clinical Psychology*, 1977, 45, 896-907.
- Rosen, R. C., & Kopel, S. A. Penile plethysmography and biofeedback in the treatment of a transvestite-exhibitionist. *Journal of Consulting and Clinical Psychology*, 1977, 45, 908-916.
- Zuckerman, M. Physiological measures of sexual arousal in the human. *Psychological Bulletin*, 1971, 75, 297-329.

Received December 2, 1977 ■

motor, and physiological responses. In our study we used all three measures to support the inference of therapeutic change. We were able to demonstrate decrements in pedophilic tendencies during and immediately after treatment and on extended follow-up, with changes evident in penile measures, self-report, and also on necessarily indirect measures of motor behavior. The last index is perhaps most important, since the ultimate criterion is whether or not the patient molests children or gets into other difficulties following treatment. At last report, our patient had not done so. Farkas (1978) implies overemphasis on the penile measure, which we feel is not the case.

A final point needs to be made regarding a possible misunderstanding on Farkas' (1978) part. In discussing the importance of treating motor behaviors, he says,

Targeting penile response without concomitant attempts to modify cognitive and behavioral factors that may cooccur or predominate as the basis for modification of sexual behaviors does not appear to hold promise for maintenance and generalization of appropriate responses in the natural situation. (p. 1516)

This implies that we used penile response as the reaction to be treated. This is not the case. In treatment we used imagined versions of real-life situations that had in the past, and might in the future, serve as occasions for sexual attraction to children. We also instructed the patient to imagine the pedophilic thoughts and fantasies that occurred in these situations.

In summary, changes in penile erection measures obtained in a laboratory setting are obviously not to be regarded as the sole criterion of therapeutic change. Such measures need to be supplemented with self-report techniques and

evaluations of motor behavior. The fact that erection can be voluntarily controlled is, indeed, a problem that may very well be almost impossible to get around (Laws & Holmen, Note 1). As Farkas himself seems to suggest, it may be best to capitalize on such control by using it in treatment instead of regarding it as an assessment problem.

Reference Note

1. Laws, D. R., & Holmen, M. L. *Sexual response faking by pedophiles*. Unpublished manuscript, Atascadero State Hospital, Atascadero, California, undated.

References

- Barlow, D. H. Assessment of sexual behavior. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Farkas, G. M. Comments on Levin et al. and Rosen and Kopel: Internal and external validity issues. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1515-1516.
- Henson, D. E., & Rubin, H. B. Voluntary control of eroticism. *Journal of Applied Behavior Analysis*, 1971, 4, 37-47.
- Levin, S. M., Barry, S. M., Gambaro, S., Wolfensohn, L., & Smith, A. Variations of covert sensitization in the treatment of pedophilic behavior: A case study. *Journal of Consulting and Clinical Psychology*, 1977, 45, 896-907.
- Rosen, R. C., & Kopel, S. A. Penile plethysmography and biofeedback in the treatment of a transvestite-exhibitionist. *Journal of Consulting and Clinical Psychology*, 1977, 45, 908-916.

Received April 20, 1978 ■

Role of Penile Tumescence Measurement in the Behavioral Treatment of Sexual Deviation: Issues of Validity

Raymond C. Rosen and Steven A. Kopel

Department of Psychiatry, College of Medicine and Dentistry of New Jersey
Rutgers Medical School, Piscataway

Farkas has raised three specific issues concerning the validity of penile tumescence assessment of sexual arousal in the laboratory. Some of these issues had been addressed in our original case study, and this reply to Farkas is intended to place the issues of validity in a broader perspective. While acknowledging the limitations of the erection measure, we nevertheless see an important place for it in both the assessment and treatment of sexual disorders.

In Rosen and Kopel (1977), we reported the results of a case study using penile tumescence measurement and biofeedback to modify the transvestite/exhibitionistic pattern in a 45-year-old male client. Despite an initial positive outcome, remission of the deviant behavior was reported to us by the client's spouse approximately 2½ years after treatment. Farkas (1978) has raised issues concerning the validity of the tumescence assessment as described in Rosen and Kopel, in addition to the broader issues of the usefulness of tumescence measurement in clinical studies. Some of the issues addressed by Farkas had, in fact, been dealt with in our original study. However, the validity issues that have been raised deserve a broader perspective than that offered by Farkas. The purpose of this reply is to examine the question of validity of tumescence measurement in greater depth.

To begin with, it is worth considering the growing appeal of penile plethysmography as a laboratory assessment measure. Rosen and Keefe (in press) have reviewed the burgeoning field of tumescence measurement and have concluded that it is a key dependent measure for assessment of male sexual arousal. Historically, the need for objective determination of sexual preference and response has led to less reliance on self-report and greater interest in direct measures of sexual arousal (Zuckerman, 1971). Especially when treating a sex offender, there are obviously good reasons for supplementing the

client's self-report with objective physiological assessment. In fact, good clinical practice dictates the use of *multiple* criterion measures for a comprehensive assessment of clinical outcome (Keefe, Kopel, & Gordon, 1978). For example, in our case study we described the use of a standardized paper-and-pencil instrument for evaluation of sexual interaction (LoPiccolo & Steger, 1974), subjective ratings of arousal to erotic stimuli, client and spouse reports of marital/sexual satisfaction, and laboratory tumescence assessments.

Regarding the validity of tumescence measurement, Farkas (1978) raises three specific issues: (a) The fact that males are able to exercise *some* degree of voluntary control of tumescence (Henson & Rubin, 1971; Rosen, 1973) is presented as a threat to the internal validity of the measure. (b) With respect to external validity, Farkas challenges the association between strain-gauge data and self-report of overt behavior. and (c) The generalization from laboratory to "real-life" stimuli is viewed as a further threat to external validity. We will deal with each of these three issues in turn.

Levin, Gambaro, and Wolfensohn (1978), in reply to Farkas (1978), have pointed out logical inconsistencies in the suggested independent assessment of voluntary control of penile response. The potential for response "faking" is just as relevant to "prebaseline assessment" as it is to subsequent measurements. Thus, although we agree that clinicians and researchers need to be concerned with voluntary control of tumescence, Farkas' suggestions do not appear to offer a real solution. Further, in our case study we examined several sources of evidence regarding the possibility of tumescence faking.

Requests for reprints should be sent to Raymond C. Rosen, Department of Psychiatry, College of Medicine and Dentistry of New Jersey, P.O. Box 101, Rutgers Medical School, Piscataway, New Jersey 08854.

From the gradual, progressive learning curve [of penile tumescence] it appears unlikely that suppression of response was due to distraction or demand characteristics . . . Furthermore, the client's subjective ratings of arousal changed more gradually than actual penile responses—further evidence against an explanation in terms of demand characteristics. Finally, penile tumescence suppression continued to improve even below the limit of sensory awareness (approximately 10% full erection). (Rosen & Kopel, 1977, pp. 914-915)

The second issue, the *external* validity of tumescence measurement (i.e., the notion of response generalization), is complex and merits further discussion. In part, this concern involves the consistency between physiological, self-report, and overt behavioral indices of sexual response. According to Farkas, the correlation between self-report and tumescence is weak. However, other researchers (e.g., Abel, Blanchard, Murphy, Becker, & Djenderedjian, in press) have found *strong* associations between these two measures. Furthermore, with respect to the relationship between laboratory control of tumescence and subsequent overt sexual behavior in the natural environment, other reports (e.g., Csillag, 1976) have shown meaningful generalization. Thus, there is independent empirical support for the external validity of the tumescence measure. However, in any given instance, external validity should not be taken for granted. Rather, clinicians or researchers should seek independent evidence for convergent validity (Campbell & Fiske, 1959) across response measures through a multiple-criterion assessment package.

In the treatment of sexual deviations, one frequently encounters a lack of consistency (i.e., divergence) across response systems. For instance, a number of exhibitionists treated in our clinic have exposed themselves in public without any evidence of concomitant erection or orgasm. In such cases it appears that the inappropriate behavior is functionally independent of genital response, and therefore tumescence measurement may not always be relevant. In the case study (Rosen & Kopel, 1977), however, erection and orgasm were critical components of the deviant pattern, and therefore were included in both assessment and treatment.

The third issue raised by Farkas (1978) is the question of stimulus generalization from laboratory to real-life stimuli. Our concern with this issue was reflected in our selection of assessment and treatment stimuli. For example, with respect to the client's transvestite pattern, we used his actual transvestite garments in assessment of sexual arousability. It is noteworthy that tumescence measurement revealed that contrary to our

expectations, fondling of these garments with active imagery failed to elicit arousal. However, subsequent videotaping of the complete transvestite-exhibitionistic script did prove to be a reliable eliciting stimulus for tumescence response. This suggests that the issue of real-life versus laboratory stimuli raised by Farkas may be an oversimplification. The videotape, although a more *contrived* stimulus in some respects, appeared to incorporate subtle but critical parameters of the real-life pattern. It appears from this case that treatment stimuli should not be selected a priori but rather should be determined by an empirically based, comprehensive assessment.

Finally, the unsuccessful long-term result reported in our case study raises the possibility that too much reliance was placed on penile plethysmography in assessment and treatment. On the contrary, after intensive review we have concluded that the *absence* of tumescence assessment beyond the initial 4-month follow-up period was a strategic error. Our reliance on the client's self-report during the subsequent 2-year period provided a series of false outcome assessments. We believe that tumescence measurement, as originally used, might have cued the therapists to the gradual return of the deviant arousal pattern, thus providing a critical opportunity for reinstatement of the treatment program.

References

- Abel, G. G., Blanchard, E. B., Murphy, W. D., Becker, J. V., & Djenderedjian, A. Two methods of measuring penile response. *Behavior Therapy*, in press.
- Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- Csillag, E. R. Modification of penile erectile response. *Journal of Behavior Therapy and Experimental Psychiatry*, 1976, 7, 573-575.
- Farkas, G. M. Comments on Levin et al. and Rosen and Kopel: Internal and external validity issues. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1515-1516.
- Henson, D. E., & Rubin, H. B. Voluntary control of eroticism. *Journal of Applied Behavior Analysis*, 1971, 4, 37-47.
- Keefe, F., Kopel, S. A., & Gordon, S. A *practical guide to behavioral assessment*. New York: Springer, 1978.
- Levin, S. M., Gambaro, S., & Wolfensohn, L. A reply to Farkas. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1517-1518.

- LoPiccolo, J., & Steger, J. C. The Sexual Interaction Inventory: A new instrument for assessment of sexual dysfunction. *Archives of Sexual Behavior*, 1974, 3, 585-595.
- Rosen, R. C. Suppression of penile tumescence by instrumental conditioning. *Psychosomatic Medicine*, 1973, 37, 479-483.
- Rosen, R. C., & Keefe, F. The measurement of penile tumescence. *Psychophysiology*, in press.
- Rosen, R. C., & Kopel, S. A. Penile plethysmography and biofeedback in the treatment of a transvestite-exhibitionist. *Journal of Consulting and Clinical Psychology*, 1977, 45, 908-916.
- Zuckerman, M. Physiological measures of sexual arousal in the human. *Psychological Bulletin*, 1971, 75, 297-329.

Received June 12, 1978 ■

A Note on Psychosomatic Factors in the Etiology of Neoplasms

Jonathan Kellerman
Division of Hematology-Oncology
Children's Hospital of Los Angeles
Los Angeles, California

Watson and Schuld attempted to study the relationship between psychopathology and subsequent development of neoplasms. Their results, which indicated no apparent connections between these variables, are limited due to methodological flaws, both relating to sample selection. The study sample, composed of psychiatric patients, was highly restricted along one of the variables studied—psychopathology—and was further confounded by uneven distribution of a potentially carcinogenic factor—alcoholism. Though there is no empirical evidence of psychological causation of neoplasms, the Watson and Schuld study is not the one to lay this issue to rest.

The recent study by Watson and Schuld (1977) concerning the relationship between psychopathology and subsequent development of malignant and benign neoplasms deserves credit for attempting to test several psychosomatic theories in a prospective manner. In this way, the authors commendably strove to eliminate methodological problems inherent in retrospective studies. Unfortunately, while dealing with one set of confounding variables, Watson and Schuld have neglected to consider another: the issue of restricted sample.

In the discussion of their results, Watson and Schuld made note of several caveats regarding their results, including possible weaknesses of the Minnesota Multiphasic Personality Inventory as a comprehensive catalogue of personality traits, premorbid differences between groups on variables not measured, and inability of such a study to address the issue of immediate premorbid loss of trauma. (To this last factor I would add that the study can say nothing *whatsoever* with regard to premorbid loss, *immediate or otherwise*.) They neglect, however, to consider the effect of using a sample of psychiatric inpatients as subjects.

In view of the specialized nature of the sample, it is not surprising that no significant dif-

ferences in psychopathology were found between groups. The basic finding of Watson and Schuld's (1977) study is that among a group of individuals for whom psychopathology had been diagnosed, the *degree* of psychopathology was not related to subsequent development of neoplasms. This is quite different from comparing groups of pathological versus nonpathological individuals and making a statement regarding psychopathology in the general (nonhospitalized) population. The use of a sample of people who represent a highly specialized subsample along the distribution of one of the major variables in the study (psychopathology) places strong limitations on the kinds of conclusions that can be drawn from Watson and Schuld's results. The only justifiable conclusions derived from the data are with regard to the degree of psychopathology within a pathological subgroup as it relates to neoplastic vulnerability.

In addition, the use of subjects suffering from alcohol-related problems injects another confound into Watson and Schuld's (1977) results in that such individuals may be more prone to specific types of malignancies, and this factor may obscure or outweigh any psychopathology-neoplasm relationship. In fact, these individuals are overrepresented by a ratio of 6:1 within the malignant group.

This is not to say that there is a great deal of prospective, objective data in support of psychopathological theories of neoplastic etiology such as that of Bahnson and Bahnson (1966), for there is not, and work is needed in this area. Such research, however, will need to carefully

Requests for reprints should be sent to Jonathan Kellerman, Division of Hematology-Oncology, Children's Hospital of Los Angeles, 4650 Sunset Boulevard, P.O. Box 54700, Los Angeles California 90054.

consider quality of sample in order to avoid the severe attenuation of results that Watson and Schuld's (1977) study falls victim to.

References

Bahnson, C. B., & Bahnson, M. B. Role of the ego defenses: Denial and repression in the etiology of

malignant neoplasms. *Annals of the New York Academy of Sciences*, 1966, 125, 827-845.

Watson, C. G., & Schuld, D. Psychosomatic factors in the etiology of neoplasms. *Journal of Consulting and Clinical Psychology*, 1977, 45, 455-461.

Received July 18, 1977 ■

Psychosomatic Etiological Factors in Neoplasms: A Response to Kellerman

Charles G. Watson
Veterans Administration Hospital
St. Cloud, Minnesota

Donald Schuld
Faribault County Human Services Center
Blue Earth, Minnesota

Kellerman argues that our use of psychiatric patients in a study designed to search for psychosomatic etiological factors in neoplasms may have led to our negative results. However, we suggest that the use of psychiatric patients increased the heterogeneity of the sample and probably enhanced, not limited, the likelihood of positive findings. He also suggests that our inclusion of alcoholics in the study may have masked real differences between neoplasm and control subjects. However, new analyses run on subsets of our malignancy and malignancy-control samples from which alcoholics were first deleted failed to support his contention. New analyses run to test for differences between the frequencies of various high-scale Minnesota Multiphasic Personality Inventory types in neoplasm and control groups also failed to support the view that neoplasm patients are qualitatively different from controls on the inventory.

In an earlier article (Watson & Schuld, 1977), we described a study comparing the diagnoses and Minnesota Multiphasic Personality Inventory (MMPI) responses of psychiatric patients who later developed neoplasms to those of psychiatric controls who did not. We found a considerably smaller number of significant differences between the two groups than would have been expected to appear on a chance basis, and we concluded that our results failed to support the increasingly popular theory that such growths are psychogenic.

Kellerman (1978) argues that our negative results may reflect the use of a "highly specialized" psychiatric sample and that we should have used a psychologically normal sample instead. He suggests that the homogeneity that can be inferred from the limitation of the sample to psychiatric patients may have attenuated relationships between neoplasm proneness and psychological dimensions. We believe, however, that our samples showed considerably more diversity in type of pathology than one would expect to

find among normals, including, as it did, schizophrenics, manic-depressives, involutional psychotics, neurotics, alcoholics, personality disorders, and brain-damaged patients. In fact, this heterogeneity may have increased the probability of significant neoplasm-personality relationships beyond that which might have been expected among normals. We would expect such relationships to be most conspicuous in samples with diverse and substantial amounts of psychopathology, and we suspect that our choice of a psychiatric ward sample enhanced—not reduced—the probability of positive results. Despite this presumed bias, our results were strikingly negative.

Kellerman (1978) also criticizes the inclusion of alcoholics in our samples. (Six appeared in our malignant neoplasm group and one among its controls.) He notes that since large quantities of alcohol may be carcinogenic, their inclusion might have obscured psychopathology-neoplasm relationships. If his reasoning is correct, it seems to us that if anything, the inclusion of alcoholics should have *increased* the probability of our finding relationships between personality types and neoplasm proneness, since more alcoholics appeared in the malignancy sample than in its controls. Despite this (possible, at least) bias toward positive results, our results were negative. Nevertheless, we have recalculated *t* tests for MMPI scales in subsamples of our malignancy and malignancy-control groups from which

This article was supported by the Veterans Administration Medical Research Service.

The authors are indebted to Teresa Kucala, Lyle Jacobs, Donald Bamber, and Duane Plemel for their contributions to the manuscript.

Requests for reprints should be sent to Charles G. Watson, Research Service, Veterans Administration Hospital, St. Cloud, Minnesota 56301.

subjects with alcohol-related diagnoses were first deleted. These analyses were based on 28 subjects, 14 from the malignancy group and 14 controls matched with them for age. None of the 13 resulting *ts* (which ranged from .06 to 1.55) run to test for differences between the MMPI scale means of the samples was significant, even at the .10 level. The view that our negative findings resulted from the physical effects of alcohol was not supported.

The Kellerman (1978) critique's emphasis on our finding that the relationships between neoplasm proneness and degree of pathology were low suggests the possibility that neoplasm risk may be related to *quality*, rather than amount, of measured psychopathology. Since we used a wide variety of (qualitatively different) dependent variables—eight diagnostic categories, 13 MMPI scales, and several hundred individual MMPI items—and found little evidence that any of these variables is related to neoplastic vulnerability, we have been skeptical about the possibility that qualitative differences exist. Nevertheless, we have run additional analyses to test for them.

To evaluate the qualitative difference hypothesis, we determined the two highest MMPI clinical scale *T* scores for each subject and compared their frequencies in our neoplasm and control samples with sign tests. Only 1 of the 20 sign tests (10 each for malignant/malignant-control and benign/benign-control comparisons) was significant at the .05 level. No significant differences were associated with malignant growths, but Hysteria (*Hy*) peaks were less common ($p =$

.02) among patients who later developed benign growths than among their controls. Since high *Hy* scores are generally positively correlated with psychosomatic tendencies, and since only 1 of the 20 differences was significant, these findings offer little support for the suggestion that our earlier procedures may have masked qualitative differences between neoplasm and control samples or the view that neoplasms are psychogenic.

Kellerman (1978) is correct in noting that our results can be generalized only to psychiatric patients. However, our negative results in psychiatric subjects, and the absence of any prospective study that has shown premorbid differences between neoplasm and appropriate control patients on either personality inventories or blind ratings of projective tests, lead us to doubt that important neoplasm-psychopathology relationships will emerge among normals, since both the quantity and diversity of psychopathology in the latter group is limited. Nevertheless, the question is ultimately an empirical one, and the issue will have to be resolved with additional studies, rather than speculations such as those offered by Kellerman and by us.

References

- Kellerman, J. A note on psychosomatic factors in the etiology of neoplasms. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1522-1523.
 Watson, C. G., & Schuld, D. Psychosomatic factors in the etiology of neoplasms. *Journal of Consulting and Clinical Psychology*, 1977, 43, 455-461.

Received March 3, 1978 ■

The Therapist-as-Fixed-Effect Fallacy in Psychotherapy Research

Colin Martindale
University of Maine

Studies of psychotherapy involve sampling two sets of subjects from two populations: patients and therapists. Conclusions about psychotherapy should thus be based on statistical evidence that results are reliable across both patients and therapists. In most published research concerning psychotherapy, no statistical evidence is provided that findings can be generalized beyond the particular sample of therapists studied. In spite of this, researchers tend to draw conclusions concerning psychotherapy and therapists in general. Analysis of variance designs that allow generalization of results across both therapists and patients are described. The serious problems with inappropriate analyses of variance—treating therapists as a fixed effect or ignoring the therapist factor altogether—are discussed. A review of recently published studies of psychotherapy reveals that most researchers have done one or the other of these inappropriate analyses.

Psychotherapy and related procedures involve two participants or sets of participants, therapists and patients. Consequently, the researcher who wishes to study psychotherapy is faced with the necessity of generalizing findings to two populations: a population of therapists and a population of patients. To do this, it is necessary to select random samples of both patients and therapists and then to perform statistical analyses that will provide assessments of reliability for both samples. However, the majority of studies on psychotherapy concentrate solely on reliability across patients. The consequence is that most of these studies tell us nothing about psychotherapy in general and a lot about whether the particular results found would probably be reproduced with a new sample of patients seen by exactly the same therapists who were used in the study. However, these studies have for the most part been presented and accepted as if they answered more than this question.

In analyses of variance of data concerning psychotherapy, therapist effects have been improperly handled by either treating the therapist factor as a fixed effect or by ignoring the therapist factor altogether. When therapist ef-

fects are handled in either of these ways, it is impossible to generalize results beyond the immediate sample of therapists used in one's study. This comment presents no original ideas or methods; it is merely a reminder to psychotherapy researchers that they tend regularly to use the analysis of variance in what is known in other contexts to be an inappropriate manner. Since this inappropriate usage can drastically inflate the possibility of Type I errors, the reminder would seem to be in order.¹

Fixed Versus Random Factors

Factors in analysis of variance are either fixed or random. In general, a factor is fixed if all treatment levels of interest are included in an experiment, whereas a factor is random if its treatment levels are drawn at random from a larger population of possible levels. It is clear that in research on psychotherapy there must be two random factors: therapists and patients.

¹ The necessity for generalization to more than one population is not unique to psychotherapy research. Clark (1973) has pointed out almost the same problems in psycholinguistic studies in which it is necessary to generalize to both populations of words and of subjects. Brunswik (1947), Hammond (1954), and Rosenthal (1966) have also dealt with the necessity of generalizing to more than one population in a variety of situations.

Requests for reprints should be sent to Colin Martindale, Department of Psychology, University of Maine, Orono, Maine 04473.

This has crucial implications for the choice of correct error terms in analysis of variance. In an analysis in which subjects are the only random factor, error terms are derived from estimates of within-cell variation. The degrees of freedom of these terms are dependent on the number of subjects. However, in analyses in which there are random factors besides subjects, this is not the case. Many error terms are not properly derived from within-cell error, and the degrees of freedom of these terms are not dependent on the number of subjects but on the number of levels of the other random factors.

Consider an experiment comparing the relative effectiveness of psychoanalysis and covert sensitization. There are several ways of designing such an experiment. In a crossed design, one randomly selects n therapists and N patients. All n therapists perform both types of treatment, and a subset of patients is assigned to each of the therapist-treatment combinations. If the therapist factor is taken as a random one, then to assess treatment effects, an F ratio is formed by dividing the treatment mean square by the Treatment \times Therapists mean square, with $df = 1$ and $n - 1$ in the case of two treatments. For one reason or another, it may be preferable to use a hierarchical design in which n therapists are nested under the two treatments. For example, n_1 therapists could be assigned to do covert sensitization, and n_2 different therapists could be assigned to do psychoanalysis. The N patients would be assigned at random to therapists as above. In this case, the correct denominator for the treatment mean square is the therapists within treatments mean square in forming an F to assess the main effect of treatments. In the case of two treatments, this denominator will have $2(n_1 - 1)$ degrees of freedom. In this analysis, one collapses over patients altogether in making the analysis. It does not matter whether each of the n therapists sees one or a thousand patients. A third possible design would involve a completely randomized one-way analysis of variance. This would necessitate random selection of n therapists and N patients (where $n = N$), random pairing of therapists and patients, and random assignment of these pairs to the two treatments. In this case the treatment effect is assessed against the within-cell mean square with $n - 2$ degrees of freedom for the case of two treatments. These are elementary considerations that are discussed in detail in any standard book on design such as Winer (1971) or Kirk (1968). The important point is that in each case, the effective upper limit on degrees of freedom is given by n rather than

by N . The number of therapists in one's study—not the number of patients—determines the degrees of freedom of F ratios and, hence, their power.

Typical Problem and Appropriate Solutions

If the effective N for a therapy study is n rather than N , and if it is difficult to obtain large numbers of therapists, then one is faced with a problem. An inherently variable event such as psychotherapy process or outcome intuitively calls for a large number of degrees of freedom if tests of hypotheses are to have much power. Where are these degrees of freedom to come from? The only fully satisfactory solution is to design one's study to include a lot of therapists. If this is not possible, the only legitimate hope is that if n must be small, one could pool. To do so, interactions of treatments with therapists or therapists-within-treatment effects must be small. In a crossed design, if the Treatment \times Therapists interaction can be shown to be zero, at a liberal α , it may be legitimate to pool the sum of squares for Therapists \times Treatments with the within-cell sum of squares to obtain a within-cell mean square estimate with a larger number of degrees of freedom for testing all effects. In a hierarchical design, it would be necessary to show that the therapists-within-treatment effect was zero. Then, this sum of squares could be pooled with the within-cell sum of squares. In either case, to avoid Type II errors (accepting the hypothesis that the effect in question is zero when it should be rejected), it is necessary to test the hypothesis that the effect to be eliminated is zero at a high value of α : Kirk (1968, p. 215) suggests $\alpha = .25$, but Winer (1971, p. 379) suggests $\alpha = .20$ or $.30$ for such tests. It is inappropriate to use $\alpha = .05$ or $.01$ for such tests, since this would inflate the probability of Type I errors in tests of treatment effects beyond the nominal α for such tests. It should be noted that many statisticians advocate a conservative "never pool" approach, and even those with a more moderate attitude advise against pooling effects when there is any a priori reason to expect that such effects might exist. Given this, pooling involving Treatment \times Therapists interactions or therapists-within-treatment effects must always be viewed as an extremely "liberal" practice.

Inappropriate Analyses

Treating the Therapist Factor as Fixed

There are several inappropriate solutions to the problem. One obvious solution is to treat

the therapist factor as fixed rather than random. In this case, the divisor in obtaining the F ratio for treatments will always be the within-cell mean square, which will have degrees of freedom primarily dependent on N rather than n . The problem with this solution is that it completely destroys the scientific value of the study, since all conclusions must be restricted to the particular therapists used in the study. (Of course, if one were deliberately carrying out a study of a specific set of therapists, as in a program evaluation, it would be quite proper to treat the therapist factor as fixed.)

Psychotherapy researchers who explicitly include a therapist factor in their analysis of variance designs do in fact tend to regard this factor as fixed. That this has serious effects can be demonstrated by a reanalysis of the data from several well-known studies. Paul (1966) investigated the efficacy of two types of therapy (desensitization and insight) and an attention-placebo control treatment. Five therapists each saw three patients in each of the three treatment conditions. Patients were assigned at random to these conditions. This was a crossed design: Therapists were crossed with treatments, and patients were nested under Therapists \times Treatments. Paul treated the therapist factor as fixed and obtained a number of F ratios for treatment effects. These F s were obtained by dividing the treatment mean square by the

Table 2

Analysis of Variance of Therapist Ratings of Global Patient Improvement from Truax et al.

Source	df	MS ^a	F ^a	F ^b
High vs. low conditions (A)	1	8.10	10.28**	3.06
Therapist within conditions (B within A)	2	2.65	3.36*	—
Role induction vs. no role induction (C)	1	3.60	4.57*	1.18
A \times C	1	.10	—	—
C \times B within A	2	3.05	3.87*	—
Error	32	.79		

^a Therapists fixed, from Truax et al. (1966, p. 397).

^b Therapists random.

* $p < .05$.

** $p < .01$.

within-cell mean square. Significant F s from the Paul study are given in Table 1. It may be asked, what would happen to the F s if we were willing to believe that therapists were selected at random? The answer is shown in the second column of Table 2. The F s here were obtained by dividing the treatment mean square by the Treatment \times Therapists mean square. Two of the F s are no longer significant ($p > .05$), and p falls from less than .01 and .001 to less than .05 for the others. Paul's original F values tell us what would probably happen if his study were replicated with exactly the same therapists and different patients. Our F values tell us what might be expected if his study were replicated with other therapists and other patients.

An example of an often-cited hierarchical study in which the therapist factor was fixed was done by Truax et al. (1966). The study concerns four psychiatrists who were rated on the therapeutic conditions of accurate empathy, genuineness, and warmth. The psychiatrists were placed into a group that offered high levels of these conditions and a group that offered low levels. Each of the therapists saw 10 patients, half of whom had previously undergone a "role induction" interview and half of whom had not. Thus, therapists were nested under levels of therapeutic conditions; therapeutic conditions and therapists within therapeutic conditions were crossed with role induction; and patients were nested under Therapists \times Role Induction. Truax et al. treated therapists as fixed and thus

Table 1

Analyses of Variance of Therapist Ratings of Six Dependent Variables From Paul (1966)

Treatment effect	Therapist fixed	Therapist random
	F ^a	F ^b
Specific improvement in performance anxiety	5.57**	6.88*
Improvement in other areas	12.04***	5.78*
Specific prognosis	3.79**	4.36
Appropriateness of type of therapy	17.55***	4.84*
Appropriateness of length of therapy	12.51***	5.39*
Therapist comfort	8.15***	4.11

^a df = 2, 30.

^b df = 2, 8.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

used the within-cell mean square to test all effects and interactions.

Results of their analysis of therapist ratings of global improvement are given in Table 2. If therapists are treated as random rather than fixed, then the correct divisor for the conditions mean square is the therapists within conditions mean square, and the correct divisor for the role induction mean square and for the Conditions \times Role Induction mean square is the Role Induction \times Therapists Within Conditions mean square. Results when the therapist factor is taken as random are shown in the last column of Table 2. As can be seen, none of the F s even approach significance. Truax et al.'s results apply only to the four therapists in their study. There is no reason to expect them to replicate.

Ignoring the Therapist Factor

Another illegitimate approach is to ignore therapists altogether in the analysis. That is, one might do a study involving a crossed or hierarchical design and then analyze it as if it were a completely randomized one-way design. This is, of course, incorrect for a variety of reasons. Not the least of these reasons is that in doing this, nonindependent observations are treated as if they were independent. The numbers in psychotherapy research are often things such as ratings of improvement by therapists. If one therapist makes more than one of these ratings, then these ratings cannot be expected to be independent. Inclusion of a therapist factor allows the researcher to take account of this. One ought to include in an analysis all factors that might reasonably be expected to have some systematic effect. It would be difficult to argue that individual therapists would have no such effect.

It is quite true that ignoring a relevant source of variation in analysis of variance will generally inflate the within-cell error term, but this is beside the point if the correct error term is not used in the first place. For the types of designs that I have been discussing, it is definitely not the case that the within-cell mean square will be inflated sufficiently to allow its use. In fact, as N becomes large, one virtually guarantees a significant treatment F even given that the treatments variance is 0 if there are any therapist differences in effectiveness.

Because ignoring the therapist factor leads to a loss of necessary error terms in an analysis in which the therapist factor is random, it involves an implicit assumption that therapists constitute a fixed factor and is inappropriate for the rea-

sons already given. Finally, ignoring the therapist factor leaves no statistical way of specifying the generality of our findings vis-à-vis the therapist population or even the sample of therapists used in a particular study.²

Discussion

Several counterarguments to the one presented here can be raised. Perhaps the most persuasive is that therapists in psychotherapy studies are not in fact chosen in a purely random manner from a defined population and that the therapist factor must therefore be conservatively considered as fixed. In reality, regarding therapists as fixed is not conservative at all, since it is generally much easier to obtain statistically significant results if the therapist factor is treated as fixed. Even if the researcher is scrupulous in pointing out that this factor is being considered as fixed and that results should not be generalized across therapists, the very fact that the results have been published strongly implies that they are believed to be of some generality.

Another objection would be that these criticisms could just as well be leveled at psychology experiments in general, that to ask psychotherapy researchers to include therapists in their designs is no more reasonable than to ask experimental psychologists to include experimenters in theirs. Actually, this is not a bad idea. Hopefully, experimentalists who use several experimenters do in fact assess the reliability of their results across experimenters either by including them in preliminary analyses or by some less

² Virtually all recent psychotherapy research has handled therapists in one or the other of these inappropriate manners. To get some idea of the pervasiveness of these design problems, the 1975 issues of the *Journal of Consulting and Clinical Psychology* and the 1973 and 1974 issues of the *Journal of Abnormal Psychology* were surveyed. Brief reports and articles with $n = 1$ were excluded. The search yielded 33 articles reporting the application of analysis of variance to results of studies of psychotherapy or related techniques such as interviewing, counseling, and behavior modification. In only 1 of these was the therapist factor treated as random. It can be inferred from reported degrees of freedom that the therapist factor was treated as fixed in 3 of the studies. Therapist effects were inappropriately ruled out (by, e.g., failing to test all effects needed for error terms or using either $\alpha < .20$ or unspecified α in preliminary tests) in 8 of the studies. The therapist factor was ignored in the analyses of variance reported in the remaining 21 articles.

formal procedure. This becomes important to the extent that the experimenter's role goes beyond that of a passive observer. Be this as it may, it is hopefully the case that experimenters are usually not explicitly included in analyses, in that their effects are negligible and their presence would merely clutter the presentation. The point is not that every conceivable factor should be included in analyses of variance but merely that factors of generally agreed-upon substantive importance should be included.

Another argument that can be quickly disposed of is that the results of psychotherapy studies do in fact replicate well. In this view, the theoretical problem of lack of generality is compensated for by the empirical fact of repeated replication. One may consult the appendix to the review by Luborsky, Chandler, Auerbach, Cohen, and Bachrach (1971) to see the fallacy of this argument. With few exceptions, there is unimpressive consistency in findings concerning the same hypothesis. Ultimately, the reliability of any phenomenon is demonstrated by repeated independent replications. The p values attached to F ratios serve to give other researchers some idea as to whether attempts at replication would probably be successful. Viewed from this perspective, p values derived from analyses of variance in which the therapist factor is fixed are grossly misleading.

A random factor is actually defined by two attributes: (a) its levels are sampled, and (b) this sampling is done at random. What does one do if levels of the therapist factor have been sampled but the sampling was not random? The conventional advice is that if levels of a factor are not sampled randomly, then the factor must be regarded as fixed. But this advice is given in the hope that if a sampled factor were so regarded, no credence to the generality of the results would be given.

Given that one must regard therapists as a random factor if a study is to have any scientific value, is there any justification for doing so? Actually, the same reasons used to treat subjects/patients as a random factor may sometimes support the treatment of therapists as a random factor. The implicit assumption that allows treatment of subjects as a random factor even if they were not randomly sampled is that random "presampling" has occurred. That is, the forces that brought the particular subjects into contiguity with the investigator were random with respect to the experiment in question. The same line of reasoning could, within reason, be applied to experimenters or therapists. I do not mean to imply that "two wrongs make a right"

or that one should not make every effort at random sampling of therapists. I have merely tried to outline a rationale for arguing that some studies analyzed with the therapist factor as fixed could legitimately be reanalyzed with this factor considered as random rather than being discounted completely.

To what extent do the criticisms discussed in this comment invalidate published psychotherapy research that has been inappropriately analyzed? Unless treatment F s can be recomputed—and this is seldom possible—in studies that have treated the therapist factor as fixed or have ignored it altogether, there is good reason to be wary of generalizations based on them. As the reanalyses described above demonstrate, treating the therapist factor as random rather than as fixed can have drastic effects on the significance of F values. Effects originally reported as significant at the .01 and .001 levels may be reduced to nonsignificant levels by this change in analysis. Given this, it is quite likely that the published research on psychotherapy contains an extremely high proportion of Type I errors. We know, then, a lot less about psychotherapy than we thought we did.

References

- Brunswick, E. *Systematic and representative design of psychological experiments*. Berkeley: University of California Press, 1947.
- Clark, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 1973, 12, 335-359.
- Hammond, K. R. Representative vs. systematic design in clinical psychology. *Psychological Bulletin*, 1954, 51, 150-159.
- Kirk, R. E. *Experimental design procedures for the behavioral sciences*. Belmont, Calif.: Brooks/Cole, 1968.
- Luborsky, L., Chandler, M., Auerbach, A. H., Cohen, J., & Bachrach, H. Factors influencing the outcome of psychotherapy: A review of quantitative research. *Psychological Bulletin*, 1971, 75, 145-185.
- Paul, G. L. *Insight vs. desensitization in psychotherapy*. Stanford, Calif.: Stanford University Press, 1966.
- Rosenthal, R. *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts, 1966.
- Truax, C. B., et al. Therapist empathy, genuineness, and warmth and patient therapeutic outcome. *Journal of Consulting Psychology*, 1966, 30, 395-401.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

Received January 26, 1977 ■

Resurrection of Uniformity Assumption Myths and the Fallacy of Statistical Absolutes in Psychotherapy Research

Gordon L. Paul and Mark H. Licht
University of Illinois at Urbana-Champaign

Martindale's assertions are criticized for failure to recognize the lack of uniformity in domains and classes of variables involved in psychotherapy research, and the need to fit design and analyses to the data obtained and the questions asked. Fallacies regarding absolute or invariant use of statistical models and techniques are discussed. Included are inappropriate statistical bases for generalization, confusion in the contribution of statistical analyses to internal versus external validity, and misapplication of statistical procedures—particularly "therapists as random effects" for most psychotherapy research. Systematic accumulation of knowledge from psychotherapy research will more likely come from adequate specification of classes of variables and focused experimental questions, combined with thoughtful design and tactics, not from rigid application of textbook statistical models to areas in which they do not fit.

Kiesler (1966) drew attention to the "uniformity assumption myths" that had historically plagued research in psychotherapy. Nearly all classes of variables of interest in psychotherapy research are more heterogeneous than our short-hand descriptive labels suggest. Failing to recognize such heterogeneity resulted in neither independent nor dependent variables being adequately considered and in overgeneralizations from inadequately specified data. Recent articles give evidence that these myths of uniformity are being resurrected (e.g., Martindale, 1978; Smith & Glass, 1977; Mariotto, in press). Martindale resurrects uniformity assumption myths through assertions that results of psychotherapy research should be generalized to psychotherapy, patients, and therapists "in general"—a focus on questions known not to be fruitful for over a decade. He further extends inappropriate assumptions of uniformity to statistical "absolutes," which are asserted as requirements for research in psychotherapy.

Complexity of Variables in the Research Enterprise

Before addressing Martindale's (1978) fallacious assumptions (space limitations prohibit

discussion of his points with which we agree), a brief reminder of the complexity of variables and the nature of the psychotherapy research enterprise may aid in orientation. These issues are developed and described in detail along with recommended designs and tactics by Paul (1969).

Other than the greater number and complexity of variables involved, the principles and methods required for psychotherapy research appear to be no different than those of any other area. Because the knowledge obtained in psychotherapy research should ultimately find its way to clinical practice, both internal and external validity become especially important (Campbell & Stanley, 1966). Internal validity refers to the degree to which plausible rival hypotheses have been ruled out within a study. Without internal validity, a study is uninterpretable. There is really only one principle for establishing internal validity—Ask a question in such a way that enough controls can be introduced to unambiguously evaluate the effects of independent variables. External validity refers to generalizability beyond the specific study. Unlike internal validity, external validity can be established only by additional investigation.

Because of the lack of uniformity of labeled clients, therapists, or treatments, any study of psychotherapy should include descriptive, measurement, or control operations for each of several detailed classes of variables to establish internal validity and provide hypotheses regard-

Requests for reprints should be sent to Gordon L. Paul, Psychological Clinic, University of Illinois, Champaign, Illinois 61820.

ing external validity—hypotheses that can later be subjected to empirical test. Within the client domain, these classes of variables include (a) the clients' distressing behaviors that are the focus of treatment; (b) the clients' relatively stable personal-social characteristics; and (c) the clients' physical-social life environment. Within the therapist domain, these classes of variables include (a) the specific therapeutic techniques through which change in problem behaviors is attempted; (b) the therapists' relatively stable personal-social characteristics; and (c) the physical-social treatment environment. The third major domain, time, serves to further specify the set of circumstances for other classes of variables and to determine the focus and nature of assessments needed within and between periods related to treatment, dependent on the questions asked.

Paul (1969) proposed the ultimate questions to be answered in clinical research (including psychotherapy) as: "What treatment, by whom, is most effective for this individual with that specific problem, under which set of circumstances, and how does it come about?" (p. 44). Although no single study can ever answer these questions, specification of the aspect of the question for which answers are sought, combined with adequate description, measurement, or control of each of the classes of variables, allows stronger internal validity and meaningful accumulation of knowledge across studies. The means of obtaining answers or partial answers then becomes a question of design and tactics.

Fallacies of Statistical Absolutes

With the above reminders of the complexity of variables in the enterprise, let us turn specifically to Martindale's (1978) more troublesome assumptions and assertions. These fall into two groups: absolutes that fail to reflect reality and absolutes that inappropriately assume uniformity of design and analyses.

Absolutes That Fail to Reflect Reality

The majority of Martindale's (1978) assumptions that fail to reflect reality are mirrored in his assertion that "to treat the therapist factor as fixed rather than random . . . completely destroys the scientific value of the study, since all conclusions must be restricted to the particular therapists used in the study" (p. 1528). Reflected in this assertion is the assumption that the only justification for generalizing is random sampling, as well as a confusion concerning the contribu-

tion of statistical analyses to internal versus external validity in an experimental study. Campbell and Stanley (1966) noted the "reluctance to accept Hume's truism that *induction or generalization is never fully justified logically*" (p. 17). Unlike random sampling in survey research in which the logic of probability statistics provides justification for extrapolation of sample findings to a defined population, the focus of randomization in experimental research is nearly always on internal validity, not representativeness of the sample. Contrary to Martindale's assumptions, calling a factor or levels of a factor "random" in an investigation of psychotherapy does not make it so. Such a practice is likely to be misleading, in that others may assume statistical justification for extrapolation on inappropriate grounds, resulting in overgeneralization of findings from less powerful analyses.

The "scientific value" of a study can only be determined by multiple criteria. Scientifically meaningful *conclusions* from a particular study are dependent on the internal validity of the experimental operations—the degree to which plausible rival hypotheses have been ruled out of cause-effect relationships in that study. The generality or external validity of findings in psychotherapy research, in practice, can seldom be more than a rational-empirical undertaking, rather than a statistical one. Hypotheses concerning the generalizability of findings are, thus, strengthened to the extent that a study provides strong internal validity and thorough measurement or description of the relevant domains and classes of variables over which generalization might be expected on the basis of knowledge obtained from other sources. Conventional interpretations in the majority of texts on statistical design point out that generalization of findings from "fixed factors" should only be within the *levels* actually included in a study. With regard to therapists as a fixed factor, this would refer to therapists with similar characteristics—not just the particular therapists involved any more than conclusions would be restricted to the particular settings, instruments, times, and so forth, involved in a specific study. However, such generalizations remain hypotheses until extensions and limitations on the conditions in which findings hold are empirically tested—a problem in experimental design and tactics, not statistical models (see Paul, 1969).

Absolutes That Inappropriately Assume Uniformity of Design and Analyses

Several of Martindale's (1978) assumptions and assertions could be appropriate for a specific

experimental question regarding psychotherapy, but they appear patently inappropriate for other questions—in fact, for most questions that investigators might rationally approach. Abstracting from Martindale's comment, these include assertions that both therapists and clients must be randomly selected for experimental inclusion; treating therapists as a fixed effect in analyses of variance drastically inflates the possibility of Type I errors; the number of therapists rather than the number of clients determines the degrees of freedom and power of statistical tests; and the only way to increase the power of hypothesis testing is to increase the number of therapists included.

Not only do the above assertions reflect the myth of uniformity of therapists and clients in general, but they also extend the myth to experimental design and analyses requirements as well. As with any other area of scientific research, investigations of psychotherapy must fit the design to the questions, or aspects, of the ultimate clinical research question asked. Asserting requirements as absolutes, irrespective of the purpose of an investigation or of the purpose of a particular set of analyses can only perpetuate a "cookbook" approach that fails to recognize that statistics are only tools to aid in decision making—They cannot replace thought or careful experimental design and controls.

Martindale (1978) provides a prime example of the outcome of such a cookbook approach in his selection of data for reanalysis from Paul (1966). He asks, "What would happen to the F s if we were willing to believe that therapists were selected at random?" (p. 1528) (a *totally* unwarranted assumption) and recomputes F s treating the therapist factor as a "random effect." In fact, such an analysis would have been *inappropriate* for the data selected *even if* the therapists had really been "randomly sampled," since the major purpose of those particular data analyses was to investigate the therapist ratings to identify possible limiting conditions on the validity of treatment effects previously evaluated by objective, external means. Maximum power within reasonable assumptions would logically call for fixed effects handling of that data, no matter how therapists came to participate in the study.

Martindale (1978) also implies that reliability of therapist contributions cannot be assessed if of therapist contributions cannot be assessed if treated as fixed effects, and that p values associated with F ratios reflect the probability of replication. Elementary considerations of experimental design and statistical inference are, in fact, discussed in most texts on the subject.

However, it seems worthwhile to note that p values are dependent on N and have no reflection on practical or scientific "significance"—They merely indicate the believeability that a difference in a particular direction was obtained. They tell nothing about the probability of replicating a difference of that size again—with the same or different subjects, whether in a fixed or a random model. The reliability and strength of effects for either fixed or random factors in analyses of variance can be examined—not by p s but by correlational analyses and such statistics as coefficient alpha (Winer, 1971) and epsilon-squared (Peters & VanVoorhis, 1940).

Conclusion

The need to abandon uniformity assumption myths is no less now than when Kiesler (1966) originally summarized them. Rather than therapists and clients in general, adequate descriptive, measurement, and control operations focused on the *specific* domains and classes of variables detailed by Paul (1969) appear necessary for establishing the internal validity of a given study and for aiding in rational generalization to practice and further investigations. The failure to replicate findings in psychotherapy research is more likely a function of inappropriate attempts to generalize without recognition of the myths of uniformity than of the error terms selected in a particular analysis. The fallacy of statistical absolutes is equally worthy of abandonment if we are to progress systematically in accumulation of knowledge. Neither a "box score" approach to determining treatment effects over studies without adequate representation of their internal validity or the internal purpose of measurement (e.g., Smith & Glass, 1977), nor absolutes in requirements of design and analyses (e.g., Martindale, 1978) add clarification—no matter how sophisticated the mathematics involved. We find ourselves in strong agreement with Myers (1972) that "in practice we expect the experimenter to use his brains as well as his F ratios to draw inferences" (p. 169).

References

- Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1966.
- Kiesler, D. J. Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin*, 1966, *65*, 110-136.
- Mariotto, M. J. Meta-analysis or meta-muddling? *Clinical Psychologist*, in press.
- Martindale, C. The therapist-as-fixed-effect fallacy

- in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1526-1530.
- Myers, J. L. *Fundamentals of experimental design* (2nd ed.). Boston: Allyn & Bacon, 1972.
- Paul, G. L. *Insight vs. desensitization in psychotherapy*. Stanford, Calif.: Stanford University Press, 1966.
- Paul, G. L. Behavior modification research: Design tactics. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- Peters, C. C., & VanVoorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
- Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 1977, 32, 752-760.
- Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

Received June 18, 1978 ■

A Question as to the Validity of the Verbal Scale IQ as a WAIS Short Form

William M. Reynolds

State University of New York at Albany

Wildman and Wildman's contention as to the validity of the Verbal Scale IQ as a Wechsler Adult Intelligence Scale short form is shown to be based on an inappropriate analysis of their data. Upon reanalysis, the claim for validity is unsupported on the basis of original criteria. Results are presented that also fail to support the validity of the Verbal IQ as a short form measure.

The development of short forms of standardized intelligence tests has been of concern to both researchers and clinicians. Due to their differential component structure, the Wechsler tests have been the subject of many dissections into the short-form status (Finch, Childress & Ollendick, 1974; Levy, 1968; Luszki, Schultz, Laywell, & Dawes, 1970; Silverstein, 1970). Recently Wildman and Wildman (1977) reported results that they contend support the validity of using the Verbal IQ as a short form of the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955). They cite three criteria, originally proposed by Resnick and Entin (1971), as necessary for the validation of a short-form test: (a) There should be a significant positive correlation between the short form and the standard test form; (b) the obtained score means between the two forms should not be statistically different; and (c) there should be a high degree of concordance between forms on the IQ classification level assigned to examinees.

Wildman and Wildman (1977) found (a) a product-moment correlation of .97 between Verbal Scale IQ and Full Scale IQ ($n = 100$), (b) a t -test value of .95 ($p > .05$) between test form means, and (c) a change in diagnosis of mental retardation in 13% of the cases. They concluded from their data that the Verbal Scale IQ meets the above stated criteria and therefore is a valid short form for the WAIS Full Scale IQ. A re-examination of their data, however, disputes this conclusion. Besides the fact that as McNemar (1949) pointed out, a spurious correlation will result when a subscore that is part of

the total score is correlated with the total score, Wildman and Wildman tested the significance between the means of the Verbal Scale IQ and the Full Scale IQ via a t test for independent samples. The inappropriate use of independent samples t tests with high positively correlated dependent samples will, as Glass and Stanley (1970) pointed out, result in an overestimation of the standard error of the differences between the means, subsequently resulting in nonsignificant differences where significant differences between the two means exist. If for related measures,

$$t = \frac{\bar{D}}{\sqrt{\frac{S_1^2 + S_2^2 - 2r_{12}S_1S_2}{N}}}$$

where \bar{D} is the mean difference, S^2 is the variance, r is the correlation coefficient, and N is the sample size, is used to reanalyze Wildman and Wildman's results, a t of 5.46 ($p < .001$) is obtained, which is highly disparate from the t value originally determined. One would therefore reject the validity of the Verbal Scale short form.

To check the possibility that the above results could in part be due to heterogeneity of Wildman and Wildman's (1977) sample (psychotics, brain damaged, and normals), WAIS results from 42 normal adults, ranging in age from 20 to 65 years with a mean age of 30.46 years ($SD = 12.70$), were analyzed. The mean Verbal Scale IQ of this group was 120.00 ($SD = 11.37$), and the mean Full Scale IQ was 117.71 ($SD = 10.61$). The IQ ranges were 87-141 for the Verbal Scale and 89-134 for the Full Scale. The obtained product-moment correlation (uncorrected) between the Verbal and Full Scale IQs was .92 ($p < .01$). A t test for correlated mea-

Requests for reprints should be sent to William M. Reynolds, Department of Educational Psychology and Statistics, Education 236, State University of New York at Albany, Albany, New York 12222.

tures between the means of the Verbal Scale and Full Scale IQs produced a t of 3.33 ($p < .01$).

The results of this study also fail to support the contention of the Verbal Scale IQ as a valid WAIS short form when nonsignificant differences between test form means are used as a criterion.

References

- Finch, A. J., Childress, W. B., & Ollendick, T. H. Efficiency of WAIS short forms. *Mental Retardation*, 1974, 12, 48-49.
- Glass, G. V., & Stanley, J. C. *Statistical methods in education and psychology*. Englewood Cliffs, N.J.: Prentice-Hall, 1970.
- Levy, P. Short-form tests: A methodological review. *Psychological Bulletin*, 1968, 69, 410-416.
- Luski, M. B., Schultz, W., Laywell, H. R., & Dawes, R. M. Long search for a short WAIS: Stop looking. *Journal of Consulting and Clinical Psychology*, 1970, 34, 425-431.
- McNemar, Q. *Psychological statistics*. New York: Wiley, 1949.
- Resnick, R. J., & Entin, A. D. Is an abbreviated form of the WISC valid for Afro-American children? *Journal of Consulting and Clinical Psychology*, 1971, 36, 97-99.
- Silverstein, A. B. Reappraisal of the validity of WAIS, WISC, and WPPSI short forms. *Journal of Consulting and Clinical Psychology*, 1970, 34, 12-14.
- Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation, 1955.
- Wildman, R. W., & Wildman, R. W., Jr. Validity of Verbal IQ as a short form of the Wechsler Adult Intelligence Scale. *Journal of Consulting and Clinical Psychology*, 1977, 45, 171-172.

Received April 26, 1977 ■

Validity of the Verbal Scale IQ as a WAIS Short Form: A Reply to Reynolds

Robert W. Wildman and Robert W. Wildman II
Central State Hospital, Milledgeville, Georgia

We were criticized because of our use of improper statistics in supporting the validity of the Wechsler Adult Intelligence Scale Verbal IQ as a short form of the Full Scale IQ. We agree with Reynolds that we erred in the choice of one statistical test. However, we contend that from a practical standpoint, the use of the Verbal IQ as a short form had considerable value as a screening device and should not be discarded because of this one criticism.

We (Wildman & Wildman, 1977) presented data interpreted as being supportive of the validity of the Wechsler Adult Intelligence Scale (WAIS) Verbal IQ as a short form of the Full Scale IQ. Resnick and Entin's (1971) three criteria for evaluating the validity of the short form were used in the 1977 study: (a) There had to be a significant positive correlation between the short form and the standard test form; (b) the obtained score means between the two forms should not be statistically different; and (c) there should be a high degree of concordance between forms on the IQ classification level assigned to examinees.

It was decided to use the t test for unmatched samples because one could make either of two assumptions—that the samples were similar or dissimilar. At the time the decision was made, we felt that we did not want to make the *a priori* assumption that our short form using the Verbal Scale would be a valid short form and predict the Full Scale IQ. We thought that this was a biased alternative to choose and neglected to think through the fact that we were using the same sample of subjects and the difference this would make statistically in testing the second criterion mentioned above. We agree with Reynolds (1978) that a matched-sample t test would have certainly been the logical one to use.

The Verbal Scale of the WAIS has a very significant positive correlation with the Full Scale IQ, and there is a high degree of agree-

ment regarding the IQ classification levels assigned to the subjects. We might say that "two out of three ain't bad." Even though the Verbal Scale short form did not meet the second criterion of Resnick and Entin (1971), from a practical standpoint the main point of a short form is to save time and to classify individuals accurately, which then makes it of value as a screening device in many clinical situations. Reynolds' data with 42 normal adults indicate a high correlation between the Verbal and Full Scale IQ. However, Reynolds fails to mention how many subjects would have been misclassified by placing them in Wechsler IQ categories.

In short, Reynolds' criticism was correct, but it seems that the Verbal Scale should be accepted as a good short form because of its demonstrated clinical value rather than discarded because it failed on one of the three criteria proposed in one article.

References

- Resnick, R. J., & Entin, A. D. Is an abbreviated form of the WISC valid for Afro-American children? *Journal of Consulting and Clinical Psychology*, 1971, 36, 97-99.
- Reynolds, W. M. A question as to the validity of the Verbal Scale IQ as a WAIS short form. *Journal of Consulting and Clinical Psychology*, 1978, 46, 1535-1536.
- Wildman, R. W., & Wildman, R. W., II. Validity of Verbal IQ as a short form of the Wechsler Adult Intelligence Scale. *Journal of Consulting and Clinical Psychology*, 1977, 45, 171-172.

Requests for reprints should be sent to Robert W. Wildman, Central State Hospital, Boone Building, Milledgeville, Georgia 31062.

Received April 3, 1978 ■

Brief Reports

Effects of Type of Social Reinforcement on the Intelligence Test Performance of Lower-Class Black Children

Francis Terrell, Jerome Taylor, and Sandra L. Terrell
University of Pittsburgh

This study examined the effects of different types of reinforcement on the performance of black males on the Wechsler Intelligence Scale for Children-Revised. After each correct response, participants were given either no reinforcement, a candy reward, traditional social reinforcement, or culturally relevant social reinforcement. Children given candy or culturally relevant social reinforcement obtained significantly higher scores than children given either no reinforcement or traditional social reinforcement.

A somewhat consistent finding is that lower-class black children perform better on cognitive tasks when given tangible rewards as opposed to social reinforcers (Schultz & Sherman, 1976). Typically, social reinforcers have consisted of such verbalizations as "good" and "fine." A problem with these studies is that little attention has been given to the type of social reinforcers used. In view of the different value system among blacks (Terrell & Taylor, Note 1), it is possible that appropriate social reinforcers have not been used in previous studies. This study explored the effects of culturally relevant reinforcers on lower-class black children's performance on the Wechsler Intelligence Scale for Children-Revised (WISC-R).

Participants were 80 lower-class second-grade black children selected from a southern urban area and randomly assigned to either a nonreinforcement condition, a candy reward condition, a social reward condition, or a culturally relevant social reward condition. All children were then individually administered the short form of the WISC-R (Silverstein, 1967) by the first author, a black PhD psychologist. In the nonreinforcement condition, the children were given no reward; in the candy reward condition, chil-

dren were given an M&M after each correct response; in the social reward condition, children were given verbal praise such as good and fine after each correct response; and in the culturally relevant condition, children were given verbal praise such as "good job, blood" and "nice job, little brother" after each correct response. Scoring was done by an experienced master's level clinician who was unfamiliar with the participants and purposes of this study.

Children in the control group obtained a mean IQ score of 81.55 ($SD = 9.88$); the social reward group obtained a mean IQ score of 84.85 ($SD = 12.17$); the tangible reinforcer group obtained a mean IQ score of 92.85 ($SD = 11.39$); and the culturally relevant group obtained a mean score of 99.15 ($SD = 10.49$). Significant differences were found among the groups, $F(3, 76) = 10.37$, $p < .01$. Using Scheffé's method of post hoc comparisons, no significant differences were found between the control and social reinforcement groups. Also, no significant differences were found between the tangible and social reinforcement groups. However, children given tangible reinforcement had significantly higher mean IQ scores than children in the control group ($p < .05$). Further, children given culturally relevant rewards obtained significantly higher IQ scores than children in the control group ($p < .01$) and children in the social reinforcement group ($p < .01$).

Thus, it is suggested that the type of social reinforcer has an important effect on black children's performance on cognitive tests. Indeed, the results suggest that the use of appropriate

This study was supported in part by a grant from the Center for Minority Group Mental Health (NIMH Grant 5R01-MH26754-02).

Requests for reprints should be sent to Francis Terrell, who is now at the Department of Psychology, Texas Christian University, Fort Worth, Texas 76129.

social reinforcers is more effective than tangible reinforcers.

Reference Note

1. Terrell, F., & Taylor, J. *The development of an inventory to measure certain aspects of black nationalist ideology*. Paper presented at the meeting of the Eastern Psychological Association, New York, April 1976.

References

- Schultz, C. B., & Sherman, R. H. Social class, development, and differences in reinforcer effectiveness. *Review of Educational Research*, 1976, 46, 25-59.
- Silverstein, A. B. Validity of WISC short forms at three age levels. *Journal of Consulting Psychology*, 1967, 31, 635-636.

Received January 21, 1977 ■

The Child With Cancer: Patterns of Communication and Denial

John J. Spinetta and Lorrie J. Maloney
San Diego State University

What role do communication and denial play in the coping efforts of children with cancer? Although no conclusions are drawn regarding cause-effect relationships, results indicate that the level of family communication about the illness, as expressed in the mother's judgment of communication, is correlated with three of the four hypothesized responses in the child. The study demonstrates the usefulness of three instruments as effective tools in measuring the child's reactions to the illness. Several questions are raised as possible avenues for further cross-sectional and longitudinal research on the communication hypothesis with the use of the instruments.

If it is true that a child with cancer as young as 6 years of age is aware of the serious nature of the illness (Bluebond-Langner, 1977; Spinetta, 1974), what does the child do with this knowledge and awareness? Does the child communicate with the parents about the prognosis in an effort to seek emotional support or does the child live in silence with the knowledge? The present study is a pilot effort to test instruments that might help clarify the issue.

To delimit the term *coping* and make it operational for the study, the following behaviors in the 6- to 10-year-old child were defined as successful attempts on the part of the child to master troublesome situations relative to the illness: a nondefensive personal posture, closeness to parental figures, happiness with oneself, and the freedom to express negative feelings within the family. Basic to this view of effective coping is our assumption that communication of a child's thoughts and concerns, both happy and painful, is both a healthier state of mental well-being and a prerequisite to family support than retaining thoughts in silence (Spinetta, 1977). It was hypothesized that a child whose family allows discussion of the illness and its prognosis would be able to cope more effectively with the illness. Specifically, the child whose family indicates open channels of communication regarding the illness would (a) score as less defensive on a scale of defensiveness (Sarason, Davidson, Lighthall, Waite, & Rue-

bush, 1960); (b) express closeness to family members, as measured by an interpersonal distance scale (Spinetta, Rigler, & Karon, 1974); (c) express happiness with self, as measured by messages to self in the Family Relations Test (Anthony & Bene, 1957); and (d) feel free to express negative feelings within the family, as measured by the Family Relations Test.

Levels of openness within the family regarding communication about the illness were measured by a questionnaire filled out by the mother. Five items, each scaled 1-4, questioned the mother's view of (a) how much the patient-child knows about the illness, (b) what kinds of questions the child asks, (c) how the parent responds to the questions, (d) what kinds of questions the siblings ask, and (e) how the parent responds to the siblings' questions. The total score of the combined categories represents the level of communication within the family, with the highest score indicating the fullest levels of communication. The study was envisioned as an effort to see whether there is a correlation among the variables; it was not intended to demonstrate a cause-effect relationship. The subjects of the study were 16 children aged 6-10 years with a diagnosis of leukemia who were being treated in outpatient clinics in three local children's facilities (Spinetta & Maloney, 1975).

Results are as follows: Using the level of communication within the family as a criterion measure regarding the illness, its prognosis, and treatment, five predictors were subjected to a multiple regression analysis, as summarized in Table 1. Three of the predictors contributed significantly to the amount of variance explained, yielding a multiple correlation of .71. Defense was the strongest predictor, followed by per-

This study was supported in part by a grant-in-aid to the first author from the San Diego State University Foundation.

Requests for reprints should be sent to John J. Spinetta, Department of Psychology, San Diego State University, San Diego, California 92182.

sonal space difference—father and negative self. These three predictors supported three of the four hypotheses.

The data were then analyzed by a multivariate analysis of variance using the same variables, with families divided into communicative (above the median, $n = 7$) or quasi-communicative (at or below the median, $n = 9$). Results are summarized in Table 2. The combination of variables significantly differentiated the communicative from the quasi-communicative families ($p < .038$). The multivariate analysis of variance supports the first three hypotheses in the predicted direction.

Although no conclusions can be drawn regarding cause-effect relationships, results indicate that the level of family communication about the illness, as expressed in the mother's judgment of communication, is related to coping strategies in the child. Families in which levels of communication about the illness are high are those families in which the children (a) exhibit a nondefensive personal posture, (b) express a long-range close relationship with the parents, and (c) express a basic satisfaction with self. Freedom to express negative feelings openly within the family was not significantly correlated with level of communication. Further studies of a longitudinal and interventional nature with the use of the above instruments are needed to test whether in fact there is a cause-effect relationship, that is, whether openness in levels of family communication regarding the illness leads to closer family ties and to healthier coping responses in the child.

Further studies from multiple sources are also needed to clarify whether in fact the mother's judgment of openness is a valid judgment of the true situation; whether there is a difference in communication patterns as the children go through subsequent relapses; whether parents become more or less communicative regarding implications of the illness as the child nears death; and, above all, whether families differ

Table 2

Multivariate Analysis of Variance for Communicative Versus Quasi-Communicative

Predictor	Communi- cative ($n = 7$)	Quasi- commu- nicative ($n = 9$)	F^a	p
Defense	10.86	14.33	3.85	.038
PSD—M ^b	7.94	5.53	2.07	.174
PSD—F ^b	11.20	6.56	2.39	.146
Negative self	2.29	2.56	9.32	.009
Negative total	24.29	26.89	.12	.732
			.38	.548

Note. PSD—M = Personal space difference—mother; PSD—F = personal space difference—father.

^a $df = 1, 13$.

^b Higher score represents desire to have parents closer.

in the level of openness that their support mechanisms can tolerate.

If one of the goals of work with children with cancer is to give them access to the intra-familial sources of support that they need most to help them in their struggle, then the present effort pointing to the relationship between levels of communication and the life-threatening child's adaptive strategies is a step toward that goal.

References

- Anthony, E. J., & Bene, E. A technique for the objective assessment of the child's family relationships. *Journal of Mental Science*, 1957, 103, 541-555.
- Bluebond-Langner, M. Meanings of death to children. In H. Feifel (Ed.), *New meanings of death*. New York: McGraw-Hill, 1977.
- Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., & Ruebush, B. K. *Anxiety in elementary school children*. New York: Wiley, 1960.
- Spinetta, J. J. The dying child's awareness of death: A review. *Psychological Bulletin*, 1974, 81, 256-260.
- Spinetta, J. J. Adjustment in children with cancer. *Journal of Pediatric Psychology*, 1977, 2, 49-51.
- Spinetta, J. J., & Maloney, L. J. Death anxiety in the out-patient leukemic child. *Pediatrics*, 1975, 56, 1034-1037.
- Spinetta, J. J., Rigler, D., & Karon, M. Personal space as a measure of a dying child's sense of isolation. *Journal of Consulting and Clinical Psychology*, 1974, 42, 751-756.

Table 1
Multiple Regression: Family Communication Patterns

Predictor	R	r	df	F
Defense	.569	-.569	1, 14	6.73
PSD—F	.709	.342	2, 13	4.22
Negative self	.716	.059	3, 12	4.12

Note. PSD—F = Personal space difference—father.

Received March 1, 1977 ■

Repression-Sensitization and Health Behavior

William F. Gayton

University of Maine at Portland-Gorham

Joseph Tavormina

University of Virginia

John E. Bassett

Shelby County Penal Farm
Memphis, Tennessee

Kenneth L. Ozmon

University of Prince Edward Island
Charlottetown, Canada

This study examined the relationship between repression-sensitization (R-S) and visits to a prison infirmary for males during a 1-year period. The main effect for the R-S dimension was significant for (a) total number of visits, (b) number of medically justified visits, and (c) number of medically unjustified visits. In each instance, sensitizers had significantly more visits than repressors. Sensitizers and repressors did not differ in the proportion of justified/unjustified visits. Differences found between repressors and sensitizers in total visits were attributed to differences in the incidence of actual illness.

Bell and Byrne (1978) have suggested that repression-sensitization (R-S) may be related to health behavior. The theoretical basis for such a relationship stems from developments in psychosomatic medicine that suggest a strong relationship between illness behavior and coping style. To the extent that the R-S dimension reflects characteristic ways of coping with stress, one might expect repressors and sensitizers to differ in terms of health behavior. Byrne, Steinberg, and Schwartz (1968) reported that sensitizers indicated a greater frequency and/or severity of illness than repressors and that male sensitizers sought medical help significantly more frequently than did male repressors. One way of accounting for these differences is to hypothesize differences in perceptions concerning illness and in the typical response to illness made by repressors and sensitizers. Sensitizers, because of their excessive use of sensitizing-type behaviors in the presence of anxiety-arousing stimuli, may have significantly higher subjective estimates of their chances of contracting an illness. This would lead to an increased likelihood of engaging in specific health behaviors. Gayton, Ozmon, Bassett, and Tavormina (1976) found that sensitizers had significantly higher perceived vulnerability-to-illness scores than did either repressors or intermediates.

An alternative hypothesis would be to assume physiological differences that lead sensitizers to be more susceptible to illness. If this were the case, the higher frequency of health visits on the part of sensitizers would be related to an increased incidence of actual disease. One way of clarifying these two hypotheses would be to divide the total number of visits to a health facility into justified visits (i.e., when medical attention was required) and unjustified visits (i.e., when no medical attention was deemed necessary). If physiological differences existed, we would expect sensitizers to have significantly more medically justified visits than repressors. If the differences were strictly at the level of perceived vulnerability, we would expect sensitizers to have significantly more unjustified visits with no differences in terms of medically justified visits.

Method

The records of all inmates who had been given the Minnesota Multiphasic Personality Inventory (MMPI) during the years 1973-1974 were taken from the files of the psychodiagnostic center of a prison for adult males in western Tennessee. The selection criteria yielded 392 protocols that were scored with templates made from the Byrne Revised Repression-Sensitization Scale (Byrne, Barry, & Nelson, 1963). Three groups of 30 subjects each—repressors, intermediates, and sensitizers—were selected from the distribution on the basis of their R-S scores. The groups did not differ in terms of

Requests for reprints should be sent to William F. Gayton, Department of Psychology, University of Maine at Portland-Gorham, 96 Falmouth Street, Portland, Maine 04103.

Table 1

Means and Standard Deviations of R-S Groups on Health Behavior Measures*

Group	Health behavior measure							
	R-S scores		Total dispensary visits		Justified visits		Unjustified visits	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Repressors	19.13	4.49	9.10	7.83	5.03	3.70	4.07	5.75
Intermediates	50.23	1.92	25.93	17.10	10.43	6.93	15.50	12.54
Sensitizers	78.50	8.50	48.13	26.56	17.90	11.68	30.23	21.53

* R-S = repression-sensitization.

age ($F < 1$) or education ($F < 1$). A measure of health behavior was obtained by recording from the prison medical records the number of sick-call visits made by each inmate in the study during a 1-year period. The number of opportunities to report voluntarily for sick call was equal for all inmates in the study. A record was also made concerning whether the prison medical officer found the inmate in need of medication and/or treatment during the dispensary visit. When this was the case, the visit was counted as medically justified. When there was no indication of medication and/or treatment, the visit was counted as medically unjustified.

Results and Discussion

Table 1 presents the means and standard deviations for the three groups on the R-S scale and the three dependent measures: (a) total number of sick-call visits; (b) number of justified sick-call visits; and (c) number of unjustified sick-call visits. A one-way analysis of variance was used to examine the relationship between each dependent variable and the R-S dimension.

Results of the three analyses indicated significant F ratios on each of the dependent measures: (a) number of sick-call visits, $F(2, 87) = 32.57$, $p < .001$; (b) number of medically justified visits, $F(2, 87) = 18.96$, $p < .001$; and (c) number of medically unjustified visits, $F(2, 87) = 23.70$, $p < .001$.

Multiple post hoc comparisons of the group means were made using the Scheffé (1953) procedure. Examination of the total number of sick-call visits revealed that sensitizers made significantly ($p < .01$) more trips to the prison dispensary than either intermediates or repressors. In turn, intermediates made significantly ($p < .01$) more trips than repressors. Examination of the frequency of justified visits revealed that

sensitizers made significantly ($p < .01$) more of these visits than did either intermediates or repressors. The latter two groups did not differ significantly from one another on this measure. Examination of the frequency of unjustified visits indicated that sensitizers made significantly ($p < .01$) more of these visits than did either intermediates or repressors. The latter two groups did not differ significantly from one another on this dimension.

The present study supports the Byrne et al. (1968) finding that male sensitizers seek medical attention significantly more frequently than do male repressors. Our previous analyses did not allow us to determine whether differences between repressors and sensitizers were due to differences in susceptibility to illness, perceived vulnerability to illness, or both. In order to help choose between the two alternatives, a post hoc analysis of variance was performed on the proportion of justified/unjustified visits. The mean proportion for repressors was 1.79; intermediates, 1.05; and sensitizers, 1.61. Results of this analysis, $F(2, 87) = .94$, ns , indicate that sensitizers and repressors do not differ in the proportion of justified/unjustified visits. This indicates that the differences found between repressors and sensitizers in total visits were most likely due to differences in actual illness.

Other possible explanations remain. Repressors may ignore or deny physical symptoms and avoid visiting a health facility even when they are actually ill. If this possibility were correct, the results of this study would be interpreted quite differently. Further research on the relationship between repression-sensitization and health behavior would seem warranted.

References

- Bell, P. A., & Byrne, D. Repression-sensitization. In H. London & J. E. Exner, Jr. (Eds.), *Dimensions of personality*. New York: Wiley, 1978.

- Byrne, D., Barry, J., & Nelson, D. Relation of the revised Repression-Sensitization Scale to measures of self-description. *Psychological Reports*, 1963, 13, 323-334.
- Byrne, D., Steinberg, M. A., & Schwartz, M. S. Relationship between repression-sensitization and physical illness. *Journal of Abnormal Psychology*, 1968, 73, 154-155.
- Gayton, W. F., Ozmon, K. L., Bassett, J. S., & Tavormina, J. Perceived vulnerability to illness and the repression-sensitization dimension. *Journal of Clinical Psychology*, 1976, 32, 616-618.
- Scheffé, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 1953, 40, 87-104.

Received March 21, 1977 ■

Relationship of Two Process Measurement Systems for Group Therapy

John E. Roe and Keith J. Edwards

Rosemead Graduate School of Professional Psychology
Biola College, LaMirada, California

Ten-minute audiotapes from 42 group therapy sessions were rated using four Hill interaction matrix variables and the Truax-Carkhuff variables of empathy, immediacy, self-exploration, and confrontation. Canonical correlation analysis suggested that the two systems converged along a dimension labeled *initiating skills*. A factor analysis suggested three underlying factors of group process labeled *Initiating Skills*, *Responding Skills*, and *Discussion Skills*. Factors 1 and 3 suggest a multidimensional structure of immediacy clarified by the Hill variables. Factor 2, with high loadings on empathy and self-exploration, identified a qualitative dimension of group process not tapped by the Hill matrix.

Research on the psychotherapeutic process has produced a proliferation of process measurement systems from many diverse perspectives. Kiesler (1973) reported more than 27 direct process measures based on observations of therapy sessions and an additional 25 indirect measures. The use of process measures has been productive in identifying factors in therapy that relate to improvement (Carkhuff & Berenson, 1967). However, theoretical integration of the results from process research has been difficult, since the studies range in disorganized fashion over the entire field of psychotherapy (Meltzoff & Kornreich, 1970). Research is needed that investigates the potential convergent validity of process measurement systems that have empirically demonstrated utility. Understanding the communality among systems of measurement is essential to integrating the theories on which they are based. The purpose of the present study was to investigate the convergent validity of two systems for assessing group process, the Hill interaction matrix (Hill, 1971) and the Truax-Carkhuff (1966) dimensions of facilitative functioning.

The group process measurement system developed by Truax and Carkhuff contains the variables of empathy, immediacy, self-exploration, and confrontation. These dimensions are well-known. Detailed definitions can be found in Truax and Carkhuff (1966). The second system of process measurement investigated was Hill's interaction matrix (Hill, 1967). As used in the present study, the Hill matrix involved a 2×2

fourfold classification modeled after the work of Lewis and Mider (1973). The first dimension defines the content style of the group interaction, with the two categories being topic-centered and member-centered styles. Interactions are classified on this dimension by asking the question, "What are the group members talking about—topics (there and then) or one another? (here and now)." The second dimension is a work-style dimension. The term *work* refers to the participation of members in the roles of patient and therapist with the goal of self-understanding. The two categories on the work-style dimension are simply *prework* and *work*.

Group process interaction segments were taken from 11 therapy groups of ministers and their spouses involved in a month-long training seminar in pastoral counseling. Groups were led by experienced counselors with graduate students as coleaders. Each group consisted of approximately 10-12 persons. Empathy, immediacy, self-exploration, and confrontation were rated using $\frac{1}{2}$ -point discriminations on 5-point scales. A modification of the Hill Interaction matrix (Form G) was used to assess the group interactions. The Hill matrix variables were evaluated by means of a 32-item questionnaire that required the raters to estimate the percentage of time spent on interactions represented by the four variables (a) topic-centered prework, (b) topic-centered work, (c) member-centered prework, and (d) member-centered work.

The unit of measurement was a 10-minute audiotape taken from the midpoint of each 90-minute session. Each group was taped during the second, fourth, sixth, and eighth sessions. A total of 42 10-minute segments were placed

Requests for reprints should be sent to John Roe, 329 Pike Courts, Martinez, California 94553.

Table 1
Canonical Variable Loadings

HIM-G				Truax-Carkhuff			
1	2	3	4	E	I	SE	C
.348	.493	-.266	.768	.206	.994	.056	.430

Note. HIM-G = Hill interaction matrix, Form G; E = empathy, I = immediacy; SE = self-exploration, and C = confrontation.

in random order on a master tape. The tapes were evaluated by four raters who were experts in the ratings of the systems used. Two raters evaluated the tapes using the Truax-Carkhuff variables. Correlations between the ratings given by the two raters were .79 (empathy), .93 (immediacy), .72 (self-exploration), and .87 (confrontation). Two other raters rated the 42 segments using the Hill matrix. The raters' quadrant scores correlated .91 (topic-centered prework), .97 (topic-centered work), .89 (member-centered prework), .94 (member-centered work). The use of independent sets of expert raters to evaluate the group interactions is considered a major strength of the present study. Communitality between the two systems of measurement was examined using canonical correlation analysis. The multivariate structure of the eight process measures was investigated using factor analysis.

Only the first canonical correlation ($r_c = .58$) was statistically significant, $\chi^2(16) = 28.5$, 16, $p = .03$. There was .339 shared variance between the two sets of variable. The first canonical variate, given in Table 1, shows the highest loadings for topic-centered work, member-centered work, immediacy, and confrontation.

The results of the factor analysis expanded the findings. Factor loadings for the first factor parallel the structure of the canonical variate with loadings on member-centered work (.86), immediacy (.63), and confrontation (.72). The second factor consisted of empathy (.91) and self-exploration (.91), and the third factor was defined by the variables topic-centered prework (.75), topic-centered work (.89), and immediacy (.51). All other loadings were below .25.

The first canonical variable suggests overlap of the two process systems on a single dimension, which could be termed *initiating skill*. The Truax and Carkhuff variables involved are action oriented, having a here-and-now emphasis that parallels Hill's concept of work in which group members assume the roles of helpers. The factor analysis sheds more light on the structure of the eight process variables as a group. Factor 1 was the interbattery factor. The second

factor encompasses the facilitative responding skills of Truax and Carkhuff. The third factor appears to be related to discussion skills. The multivariate nature of the relationships among these process measures supports Carkhuff's view that effective interpersonal process involves both responding and initiating. It was expected that empathy and self-exploration would be more highly correlated with member-centered work than they were, based on Hill's position that Quadrant 4 has the highest therapeutic value. However, the results show an independence of these variables. The important concepts of empathy and self-exploration are not accounted for by Hill's variables. It appears that while Hill's Quadrant 4 identifies the conceptually important here-and-now character of group process, assessment of whether the interaction is facilitating or detracting is not possible.

When considered together, the two process systems complement and enhance each other. Hill's system shows that immediacy as conceptualized by Truax and Carkhuff is multidimensional, involving both topic-centered and member-centered interactions. The empathy and self-exploration variables of Truax and Carkhuff suggest a qualitative dimension not explicit in the Hill schema. Further research on the development of a system that incorporates both aspects holds promise for enhancing our understanding of group therapy.

References

- Carkhuff, R. R., & Berenson, B. G. *Beyond counseling and therapy*. New York: Holt, Rinehart & Winston, 1967.
- Hill, W. F. The Hill interaction matrix. *Personnel and Guidance Journal*, 1971, 40, 619-622.
- Kiesler, D. J. *The process of psychotherapy*. Chicago: Aldine, 1973.
- Lewis, J., & Mider, P. A. Effects of leadership style on content and work. *Journal of Counseling Psychology*, 1973, 20, 137-141.
- Meltzoff, J., & Kornreich, M. *Research in psychotherapy*. New York: Atherton Press, 1970.
- Truax, C. B., & Carkhuff, R. R. *Toward effective counseling and psychotherapy*. Chicago: Aldine, 1966.

Received June 10, 1977 ■

Development and Validation of a Heterosocial Skills Inventory: The Survey of Heterosexual Interactions for Females

Carolyn L. Williams and Anthony R. Ciminero
University of Georgia

A self-report heterosocial skills inventory for females (SHI-F) was developed and was found to have satisfactory test-retest reliability, excellent internal consistency, and significant correlations with self-reported assertiveness and anxiety measures. An initial validity study compared high and low SHI-F scorers in analogue situations in which self-report, behavioral, and heart rate measures were taken. Although heart rate did not differ between groups, some behavioral and all self-report differences were significant. Additional data suggest that the SHI-F is more a measure of social skills and general negative self-evaluations and less a measure of interpersonal anxiety.

Recently there has been a growing interest in including persons who have difficulties in heterosocial situations as subjects in analogue treatment studies. Since the difficulties are thought to be characteristic of "minimal daters," college students who report infrequent dating have often served as subjects. Studies on minimal daters have shown their problems to be frequently occurring (Martinson & Zerface, 1970), clinically relevant (Curran, 1975; Twentymen & McFall, 1975), and not ameliorated by demands or suggestions for improvement (Borkovec, Stone, O'Brien, & Kaloupek, 1974). However, most of the studies in this area have been primarily concerned with treatment techniques and have neglected subject selection and assessment procedures. In addition, most of these studies have focused exclusively on male undergraduates, thus neglecting the female population.

The present study developed and initially validated an instrument for females based on Twentymen and McFall's (1975) Survey of Heterosexual Interactions (SHI). The Survey of Heterosexual Interactions for Females (SHI-F) was made as similar as possible to the SHI. Each survey contains four questions on dating frequency and 20 heterosocial situations in which the subjects are requested to rate on a 5-point scale their ability to initiate or carry on a conversation in that situation. An item on requesting the subject to rate her physical at-

tractiveness was added to the SHI-F. A low score on the SHI-F indicates a less heterosocially skilled individual.

The first part of this study collected normative and reliability data and described the characteristics of subjects selected by the SHI-F. The SHI-F was administered to 256 undergraduate females in introductory psychology classes at the University of Georgia. The mean score on the SHI-F was 68.28, with scores ranging from 32 to 98. The standard deviation was 11.78. The internal consistency of the SHI-F was substantially high as measured by the coefficient alpha ($\alpha = .89$), and its test-retest reliability was acceptable, $r(38) = .62$, $p < .001$. High and low scorers (at least 1 standard deviation above or below the mean) on the SHI-F did not differ on questions requesting them to estimate their average number of dates. However, high scorers reported dating a significantly greater number of different males per year, $t(73) = 2.75$, $p < .01$, rated themselves as participating in a greater amount of heterosocial behavior, $t(72) = 3.35$, $p < .01$, and rated themselves as significantly more attractive, $t(71) = 3.62$, $p < .001$, than the low scorers. In addition, the SHI-F was significantly correlated with the Rathus Assertiveness Scale, $r(119) = .558$, $p < .001$, and the Trait portion of the State-Trait Anxiety Inventory, $r(117) = -.404$, $p < .001$.

From the initial 256 subjects, a group of 15 high scoring and a group of 15 low scoring subjects were compared in six heterosocial situations in which physiological, behavioral, and self-report measures were collected.

The heterosocial situations were adapted from Rehm and Marston (1968) and required each

Requests for reprints and for an extended report of this study should be sent to Carolyn L. Williams, Department of Psychology, University of Georgia, Athens, Georgia 30602.

subject to interact over an intercom with a male confederate who was in an adjacent room. An example of the situations is as follows: "You run into a guy you dated a few times in high school in a drugstore near campus. He says 'Hi, I didn't know you were going to school here.'"

Subjects were asked to respond to the male as if they were actually in the situations described. During each situation the subject's heart rate was monitored via biotelemetry equipment. No significant differences between the two groups were found on heart rate. This contrasts with Twentyman and McFall's (1975) study, in which some analyses of the heart rate data revealed significant differences between males scoring high and low on the SHI. For females, the physiological response channel did not differentiate between the two groups.

Several behavioral measures were also taken during the six social behavior situations. Two observers rated the subject's overall anxiety and social skill plus three specific anxiety indicators and three specific social skill indicators during each scene. They also rated the subject's physical attractiveness after the last situation. Two time measures, duration and latency of response, were also recorded for each scene. (Interrater reliabilities for the behavioral and physiological measures are presented in the extended report.)

Significant differences were found on observer-rated social skill, with the high group being rated as more socially skilled, $t(28) = 2.17$, $p < .05$. Two of the three specific social skill indicators, interest and initiation, also differentiated between the two groups, with subjects in the high group rated by observers as showing more interest, $t(28) = 2.29$, $p < .05$, and more initiation, $t(28) = 1.95$, $p < .05$. The groups were not rated as different on the overall anxiety measure, on any of the specific anxiety indicators, on physical attractiveness, or on either of the time measures.

The subjects also rated their own overall anxiety and social skill in each scene. In contrast to the behavioral measures, the two groups differed on both social skill, $t(28) = 4.36$, $p < .001$, and anxiety, $t(28) = 4.89$, $p < .001$, in the situations. The low group subjects also rated

themselves as less attractive than the high group subjects, $t(28) = 3.03$, $p < .01$. Thus, it appears that low scoring subjects on the SHI-F perceive themselves and their behavior as less adequate than do independent observers.

These results point to the necessity of a multiple channel approach when studying a new self-report device and assessing the differences between groups. If only the self-report channel had been used in the present study, it would have appeared that the SHI-F measured both heterosocial skill and anxiety. However, the additional information from the behavioral channel suggested that the SHI-F was a heterosocial skills inventory.

The results of this study demonstrated the SHI-F to be a reliable and potentially useful instrument for selection of female subjects for analogue research in the heterosocial skills area. Use of the SHI-F in this manner would ensure more precise specification of subjects and more standardization across studies with females. Further research is needed to examine additional psychometric properties of the SHI-F and its use as an outcome measure in treatment studies.

References

- Borkovec, T. D., Stone, N. M., O'Brien, G. T., & Kaloupek, D. G. Evaluation of a clinically relevant target behavior for analogue outcome research. *Behavior Therapy*, 1974, 5, 504-514.
- Curran, J. P. Social skills training and systematic desensitization in reducing dating anxiety. *Behaviour Research and Therapy*, 1975, 13, 65-68.
- Martinson, W. D., & Zerface, J. P. Comparison of individual counseling and a social program with nondaters. *Journal of Counseling Psychology*, 1970, 17, 36-40.
- Rehm, L. P., & Marston, A. R. Reduction of social anxiety through modification of self-reinforcement: An instigation therapy technique. *Journal of Consulting and Clinical Psychology*, 1968, 32, 565-574.
- Twentyman, C. T., & McFall, R. M. Behavioral training of social skills in shy males. *Journal of Consulting and Clinical Psychology*, 1975, 43, 384-395.

Received June 20, 1977 ■

Interpersonal Liking and Self-Disclosure

Richard Gelman and Hugh McGinley
University of Wyoming

Sixty-six female subjects viewed a videorecording of a female stranger who discussed her opinions about 10 social issues. After this, the subjects rated the stranger on the Interpersonal Judgment Scale and indicated on the Jourard Self-Disclosure Questionnaire what topics they would be willing to discuss with the stranger. What the subjects would discuss was found to be positively related to their characteristic level of disclosure and their attraction toward the stranger. The implications of the results for therapist-client interactions are discussed.

Self-disclosure can be viewed as a process in which an individual purposely communicates information about himself or herself. It is a topic that in the last several years, has become the focus of considerable research activity. Partially, this activity can be related to the assumed importance of self-disclosure to mental health. Some therapists and counselors believe that it is vital to have knowledge of significant aspects of their clients' lives if they are to work effectively with them. Inasmuch, a clear understanding of the effect of their own self-disclosure on the clients' self-disclosure would be useful to them.

Perhaps the strongest finding in the clinical analogue study of self-disclosure is that individuals are willing to disclose more about themselves to others whom they like than to others whom they dislike or regard with indifference (Fitzgerald, 1963; Halvorsen & Shore, 1969; Jourard & Lasakow, 1958; Worthy, Gary, & Kahn, 1969). Based on the relationship between self-disclosure and liking, it seems reasonable to expect variables that affect interpersonal liking to also influence self-disclosure. A variable that has been unequivocally shown to influence interpersonal liking is attitude similarity.

The assumption is that attitude similarity is a positive reinforcement (Byrne, 1971) and that people are interpersonally attracted to others who are associated with the positive emotional states that are elicited by reward, the similar attitude (Lott & Lott, 1968). Knecht, Lippman, and Swap (1973) have used attitude similarity as a manipulation of interpersonal

liking in a self-disclosure study. In this study, subjects perused an attitude questionnaire that had been purportedly filled out by a stranger who was either similar or dissimilar to the subjects in attitude. After their perusal of the questionnaire, the subjects rated the stranger on the Interpersonal Judgment Scale (IJS) and then indicated what they would be willing to discuss with the stranger. Knecht et al. found that subjects were willing to discuss more items with a stranger who held similar attitudes as compared to dissimilar attitudes. However, based on trends in their data, Knecht et al. suggested that it was attraction (liking) toward the stranger that determined the subjects' willingness to disclose, and not attitude similarity *per se*.

The purpose of the present study was to further investigate the relationship between liking, attitude similarity, and disclosure to a stranger.

Method

The subjects were 66 females from introductory psychology classes who had, 6-8 weeks earlier, completed a 36-item attitude survey. The materials of the study were four videotapes of a female stranger expressing her views on 10 of the items from the 36-item questionnaire, a Sony Videocorder, a 21-inch television monitor, the 40-item Jourard Self Disclosure Questionnaire (JSDQ; Jourard & Resnick, 1970), and the IJS (Byrne, 1971).

The subjects met in coacting groups of 2-8 persons. They first completed the JSDQ by indicating which items they would be willing to discuss with a female stranger of their own age. Next they viewed one of the four videotapes of a stranger talking about her attitudes, and then

Requests for reprints should be sent to Hugh McGinley, Department of Psychology, University of Wyoming, Laramie, Wyoming 82071.

they rated the stranger on the IJS. Finally, the subjects indicated on the JSDQ what they would be willing to discuss about themselves with the stranger. The stranger expressed different viewpoints on each of the four videotapes. This was done to ensure a wide range of attitude similarity-dissimilarity between the stranger and the subjects. This procedure is a modification of Byrne's (1971) standard stranger technique.

The score compiled for the JSDQ was the sum of the items the subjects indicated that they would be willing to discuss. These scores could range from 0 to 40. The IJS score was the sum of two items, liking and desirability as a work partner. This score could range from 2 to 14, with the lower score being positive. The attitude similarity-dissimilarity score was based on the differences between the stranger's expressed attitudes about the 10 topics that she discussed and the subjects' responses to these same 10 items on the attitude questionnaire. Each item offered eight possible answers ranging from "absolutely in favor" to "absolutely opposed." The subject's response to a given item was scored as zero if it was within one scale unit of the stranger's response, and it received a difference score of one for each unit of discrepancy thereafter, regardless of direction. The theoretical range of the attitude similarity score was 0-60, with a low score indicating attitude similarity.

Results and Discussion

The purpose of the first analysis was to discern which of the three variables, Disclosure 1 (the subject's first completion of the JSDQ), attitude similarity, or interpersonal liking (the IJS score), would best predict the subject's level of disclosure to the stranger (Disclosure 2). The regression equation for these data was $\text{Disclosure 2} = 9.75 + .77 \text{ Disclosure 1} - .94 \text{ Liking} - .08 \text{ Attitude Similarity}$. The multiple correlation coefficient was .77, $F(3, 62) = 30.06$, $p < .001$. The next analysis described the individual relationships between Disclosure 2 and the three predictor variables.

The standard regression method of decomposition was used (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975). In this method, each predictor variable is treated as if it had been added to the regression equation in a separate step after the other two predictors had been included. The amount of Self-Disclosure 2 variance explained by Disclosure 1 was significant, $F(1, 62) = 82.86$, $p < .001$, as was that explained by interpersonal liking, $F(1, 62) = 6.73$,

$p < .025$, whereas attitude similarity was not a significant predictor ($F < 1$).

As expected, the best predictor of the subject's disclosure to the stranger was her score on the first administration of the JSDQ. Beyond this, disclosure to the stranger was best predicted by her rating of the stranger on the IJS. This suggests that although self-disclosure may be an enduring trait, it is influenced by interpersonal liking, which is determined, in this study at least, by a variable or variables other than attitude similarity. This result raises a question. According to Byrne (1971), interpersonal attraction is directly related to the degree of attitude similarity between two people. Indeed, in the present study, the correlation between these two variables was .73 ($df = 64$, $p < .001$).

In the decomposition of the multiple regression, the liking score was significantly related to the subject's disclosure to the stranger, but attitude similarity was not. This suggests that although attitude similarity was a contributor to the liking score, its contribution was not significantly related to the subject's disclosure to the stranger. There was at least one other systematic contributor to the IJS score.

It is possible that there was an implicit attitude similarity that influenced the subject's attraction toward the stranger. The stranger's behavior of disclosing her attitudes about such topics as premarital sex and God implied both that she was willing to disclose and that she had a positive attitude toward disclosure. In the same vein, it can be said that a subject with a higher disclosure score would also have a positive attitude toward disclosure. Thus, a high disclosing subject shared a common and positive attitude with the stranger, whereas a low disclosing subject did not. Following from this, we would expect a high discloser, regardless of attitude similarity on other topics, to be more attracted to the stranger than a low discloser would be.

To test this possibility, the 66 subjects were trichotomized on the basis of their initial JSDQ scores, and the IJS scores made by the two extreme groups were compared. The high disclosers were significantly more positive toward the stranger than were the low disclosers, $t(42) = 3.15$, $p < .01$. Thus, similarity in self-disclosure was an important factor in determining the subject's attraction toward the stranger, and it was, apparently, this source of interpersonal liking that influenced the subject's disclosure to the stranger.

The results indicate that people are more attracted to others whose self-disclosure is simi-

lar to their own level of disclosure and that this attraction has an effect on the other's disclosure. This finding implies that the congruity of disclosure level between therapist and client may not be as crucial for the client whose characteristic level of disclosure is high, since the client is likely to continue her or his personal disclosure in the presence of the therapist. However, if a client who has a characteristically low level of disclosure perceives her or his therapist to be overly disclosing, then there is the possibility that the client might develop negative feelings toward the therapist that could preclude or interfere with successful counseling.

References

- Byrne, D. *The attraction paradigm*. New York: Academic Press, 1971.
- Fitzgerald, M. P. Self-disclosure and expressed self esteem, social distance, and areas of self-revealed. *Journal of Psychology*, 1963, 56, 405-413.
- Halvorsen, C. F., Jr., & Shore, R. E. Self-disclosure and interpersonal functioning. *Journal of Consulting and Clinical Psychology*, 1969, 33, 213-217.
- Jourard, S. M., & Lasakow, P. Some factors in self-disclosure. *Journal of Abnormal and Social Psychology*, 1958, 56, 91-98.
- Jourard, S. M., & Resnick, J. L. Some effects of self-disclosure among college women. *Journal of Humanistic Psychology*, 1970, 10, 84-93.
- Knecht, L., Lippman, D., & Swap, W. Similarity, attraction, and self-disclosure. *Proceedings of the 81st Annual Convention of the American Psychological Association*, 1973, 8, 205-206. (Summary)
- Lott, A. J., & Lott, B. E. A learning theory approach to interpersonal attitudes. In A. G. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), *Psychological foundation of attitudes*. New York: Academic Press, 1968.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. *Statistical package for the social sciences*. New York: McGraw-Hill, 1975.
- Worthy, M., Gary, A. L., & Kahn, G. M. Self-disclosure as an exchange process. *Journal of Personality and Social Psychology*, 1969, 13, 59-63.

Received June 27, 1977 ■

Reported Stressful Events During Developmental Periods and Their Relation to Locus of Control Orientation in College Students

Stephen Nowicki, Jr.
Emory University

To ascertain whether high levels of stress at different periods of development may be related to an external locus of control, 30 externals (15 males and 15 females) and 30 internals (15 males and 15 females) completed the Life Events Scale. Analysis of the data indicated that for females, stress in the preschool and pubescent, and for males in the elementary and pubescent, years was related to externality. It is suggested that there may be different critical developmental periods for males and females during which high levels of stress may be related to an external locus of control.

Most authors studying antecedents of locus of control assumed that children develop a locus of control orientation by reacting to continuous parental or environmental stimuli (see MacDonald, 1971). However, Bryant and Trockel (1976) have emphasized the possibility that noncontinuous events may determine whether a child becomes internal or external in orientation. Using Adler's assumption that "individuals particularly try to make sense of their stressful or perceived unusual life experiences" (cited in Bryant & Trockel, 1976, p. 266), Bryant and Trockel hypothesized that certain discrete stressful early life experiences may predispose the individual to an external orientation. To test such a proposition, these investigators administered a life stress questionnaire (Coddington, 1972) to females that covered four time periods: preschool, elementary, junior high school, and high school. As predicted, they found that females who expressed an external locus of control on the Adult Nowicki-Strickland Locus of Control Scale reported more stressful experiences during their preschool years than did the internal females. However, they neither replicated their results for females nor generalized them to males. If their findings could be confirmed for females *and* males, they would suggest strongly that the significant antecedents of locus of control orientation may occur before a child begins school and as a result of specific events as well as continuous interactions.

Requests for reprints and for an extended report of this study should be sent to Stephen Nowicki, Jr., Department of Psychology, Emory University, Atlanta, Georgia 30322.

The subjects for the present study were the top and bottom 30% of a group of subjects ($N = 103$) whose median Nowicki-Strickland Locus of Control score of 9 was comparable to that of Bryant and Trockel (1976). These subjects (15 female internals, 15 male internals, 15 female externals, and 15 male externals) completed Coddington's (1972) Life Events Scale. In Coddington's scale, the birth of a sibling was arbitrarily chosen as a midpoint and given a score of 50. Other events receive more or less than this arbitrary midpoint value.

Analysis of the data via a 2 (male vs. female) $\times 2$ (internal vs. external) $\times 4$ (preschool, elementary, junior high school, and senior high school) analysis of variance with repeated measures on the last factor revealed a significant three-way interaction, $F(3, 168) = 3.04$, $p < .05$. Subsequent post hoc testing via Newman-Keuls procedures suggested that as predicted, external females did report more stress during their preschool years than did internal females. However, further analysis also revealed that external females reported more stress during their junior high school years than did internal females, and external males reported more stress than internal males during their elementary and junior high school years.

It appears from these results that females and males do *not* share similar histories of stressful experiences. For females, it seems that the preschool, and for males the elementary, years may be differentially significant for the development of an external orientation.

However, it appears that events occurring at the tumultuous time of pubescence may be important for how much control males *and* females perceive that they have over their life. More

work with children at each of the three levels could substantiate whether stress or the locus of control orientation comes first. Males and females seem to have different vulnerabilities to stress and to the ability of that stress to affect their locus of control orientation. Further work needs to delineate the source of those sex differences as a function of time periods.

References

Bryant, B. K., & Trockel, J. F. Personal history of psychological stress related to locus of control

orientation among college women. *Journal of Consulting and Clinical Psychology*, 1976, 44, 266-271.

Coddington, R. P. The significance of life events as etiological factors in the disease of children. *Journal of Psychosomatic Research*, 1972, 16, 7-18.

MacDonald, A. P. Internal-external locus of control: Parental antecedents. *Journal of Consulting and Clinical Psychology*, 1971, 37, 141-147.

Received July 18, 1977 ■

An Index of Premorbid Intelligence

Robert S. Wilson, Gerald Rosenbaum, Gregory Brown,
Daniel Rourke, and Douglas Whitman
Wayne State University

James Grisell
Lafayette Clinic, Detroit, Michigan

The aim of this study was to develop a way of estimating Wechsler IQs from demographic measures. The three summary IQs of the 1955 WAIS standardization sample were regressed on age, sex, race, education, and occupation. The resulting R^2 s were .53, .42, and .54 for the Verbal IQ, Performance IQ, and Full Scale IQ, respectively. The three regression equations provide actuarial indices of IQ that can be used as estimates of premorbid ability in neuropsychological assessment and research.

In neuropsychological assessment and research, there are a number of situations in which knowledge of subjects' premorbid IQ is desirable. Test data from the period preceding disease onset are, however, rarely available. Thus, clinicians are forced to estimate.

These estimates have relied on one of two data bases: (a) present ability measures thought to be relatively insensitive to neurologic disease (e.g., vocabulary) or (b) demographic information known to be related to IQ (e.g., education).

Although present ability measures such as vocabulary are highly correlated with IQ, their insensitivity to central nervous system disease is questionable. When compared with age-matched controls, neurologic patients in fact show a significant decline on all Wechsler Adult Intelligence Scale (WAIS) subtests (Russell, 1972). Impairment indices based on this model of estimating premorbid IQ have been ineffective in identifying cases of deterioration (e.g., Wechsler, 1958; Yates, 1956).

There have been two attempts to estimate premorbid IQ from demographic measures (Fogel, 1964; Ladd, 1964). The aim of both studies was to differentiate neurologic patients from controls. Subjects were assigned to one of four

educational categories. Control group means on several WAIS variables were then used as premorbid IQ estimates for patients with equivalent educational backgrounds. Use of these estimates improved Ladd's classification accuracy by 4%. Fogel did not report on the utility of the estimates.

These results are not dramatic, but they do suggest that the demographic approach may have merit, particularly given the reliance in these studies on a single demographic variable, education. Since adult onset disease should have little effect on demographic status, the accuracy of such estimates should be limited only by the correlation between IQ and the demographics. We assumed that this correlation could be substantially increased by the addition of other demographics known to be related to IQ (e.g., race, occupation). Our aim was, therefore, to develop multiple regression equations, based on a large, representative sample, that would permit estimation of premorbid IQ from demographic status.

Subjects ($N = 1,700$) consisted of the 1955 WAIS standardization sample minus the Kansas City elderly subjects (Wechsler, 1955, 1958). The three WAIS summary IQs (Verbal, Performance, and Full Scale) were regressed in a stepwise fashion on five demographic variables (age, sex, race, education, and occupation). Age and education were treated as continuous variables, whereas sex, race, and the 13 U.S. census occupation categories used by Wechsler were dummy coded (Cohen, 1968).

To simplify the equations, the unstandardized regression coefficients for the 12 dummy occupation categories were averaged across the three regression runs to yield a composite raw score

This article is based on a dissertation submitted by the first author under the direction of the second author.

Requests for reprints should be sent to Robert S. Wilson, who is now at the Department of Psychology and Social Sciences, Rush Presbyterian-St. Luke's Medical Center, 1753 West Congress Parkway, Chicago, Illinois 60612.

weight for each occupation category. The three IQs were regressed a second time on the demographic variables using these composite weights as occupation scores.

Education and race were the most powerful predictors in each equation. The r^2 s between education and Verbal IQ, Performance IQ, and Full Scale IQ were .44, .31, and .43, respectively. The R^2 s between all five demographics and Verbal IQ, Performance IQ, and Full Scale IQ were .53, .42, and .54, respectively. Thus, the four additional variables increased the amount of explained IQ variance by about 10%. With the exception of the negligible relation between sex and performance IQ, addition of each demographic variable significantly ($p < .01$) increased R^2 .

The regression equations that can be used to estimate premorbid IQ are as follows:

Estimated Verbal IQ =

$$(.18) \text{ Age} - (.02) \text{ Sex} - (.899)$$

$$\text{Race} + (3.09) \text{ Education} +$$

$$(.97) \text{ Occupation} + 70.80.$$

Estimated Performance IQ =

$$(.14) \text{ Age} - (.66) \text{ Sex} - (12.91)$$

$$\text{Race} + (2.44) \text{ Education} + (.91)$$

$$\text{Occupation} + 81.55.$$

Estimated Full Scale IQ =

$$(.17) \text{ Age} - (1.53) \text{ Sex} -$$

$$(11.33) \text{ Race} + (2.97)$$

$$\text{Education} + (1.01) \text{ Occupation} + 74.05.$$

In these equations, male = 1, female = 2, white = 1, nonwhite = 2, and the scores for Wechsler's (1955, p. 7) 13 occupation categories are 5, 1, 7, 7, 6, 3, 3, 5, 0, 1, 4, 10, and 0, respectively. The standard errors of estimate for the three equations are 10.2, 11.4, and 10.2, respectively.

These equations are based on 1955 data. The present applicability of the equations is an empirical question. If the formulae do permit

meaningful estimates of premorbid IQ, their use should aid in the identification of persons who have deteriorated intellectually. In a subsequent article we will show that these formulas can increase the accuracy with which the WAIS classifies persons as deteriorated versus normal by more than 10%.

Given the increase in educational attainment over the past 20 years, from a mean of 10.1 in the WAIS sample to a median of 12.3 in 1975 (U.S. Bureau of the Census, 1976), these equations can be expected to overestimate premorbid IQ. A partial solution to this problem consists of adjusting (deflating) the education weights in the equations to their 1955 level by multiplying the weights by .82 (10.1/12.3).

References

- Cohen, J. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 1968, 70, 426-443.
- Fogel, M. The intelligence quotient as an index of brain damage. *American Journal of Orthopsychiatry*, 1964, 34, 555-562.
- Ladd, C. WAIS performances of brain damaged and neurotic patients. *Journal of Clinical Psychology*, 1964, 20, 114-117.
- Russell, E. WAIS factor analysis with brain-damaged subjects using criterion measures. *Journal of Consulting and Clinical Psychology*, 1972, 39, 133-139.
- U.S. Bureau of the Census. *Current population reports, Educational attainment in the United States: March 1975*. (Series P-20, No. 295) Washington, D.C.: U.S. Government Printing Office, 1976.
- Wechsler, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation, 1955.
- Wechsler, D. *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore: Williams & Wilkins, 1958.
- Yates, A. The use of vocabulary in the measurement of intellectual deterioration—A review. *Journal of Mental Science*, 1956, 102, 400-440.

Received July 18, 1977 ■

Relationships Among Marital Assessment Procedures: A Correlational Study

Gayla Margolin

University of California, Santa Barbara

This study examined the relationships among three methods for assessing marital adjustment: self-reports of marital satisfaction, spouse reports of pleasing and displeasing behaviors, and trained observers' coding of positive and negative communication behaviors. Frequency of pleasing behaviors was the only measure that correlated with global marital satisfaction. Inverse relationships were found between positive and negative scores for two of the three methods investigated.

Recent investigations into the quality of marital relationships have resulted in multiple procedures to measure relationship functioning. These procedures span a variety of observers (e.g., self, spouse, and trained others), settings (e.g., home and laboratory), targets (e.g., communication, companionship, sex, and affection), and methods (e.g., global ratings, daily observations, multibehavioral coding systems) (Weiss & Margolin, 1977). The purpose of this study was to describe the relationships among several of the most commonly used assessment procedures.

A social-learning formulation of marital adjustment suggests that marital dysfunction is related to (a) a disproportionately high exchange of displeasing, as compared to pleasing, events between partners and (b) reliance on coercive, rather than constructive, methods to bring about change in problem areas. To examine these facets of relationship functioning, this study used assessment methodologies that measure couples' (a) daily exchanges of pleasing and displeasing behaviors, (b) positive and negative communication behaviors, and (c) global impressions of marital adjustment. Each assessment methodology contained a positive and a negative dimension. On the assumption that spouse behaviors are related to marital adjustment, high correlations were expected between each behavioral measure and self-reported marital satisfaction.

The 27 couples who participated in this study were self-referred for marital counseling. Their average length of marriage was 10.8 years; their

mean spouse score on the Locke-Wallace inventory was 71.8. Couples were in the pretreatment assessment phase of their therapy at the time of this study. As part of this assessment, couples (a) completed a packet of self-report inventories, (b) came into the laboratory to videotape two 10-minute negotiation discussions, and (c) kept daily records of the pleasing and displeasing behaviors received from the partner during a 1-week period.

Two self-report measures of relationship satisfaction contained in the initial assessment packets were the Locke-Wallace Marital Adjustment Inventory, a traditional measure of marital adjustment, and the Areas of Change Questionnaire, which assesses the changes that each spouse desires for self and spouse in 34 common problem areas (Birchler & Webb, 1977). Daily frequencies of pleasing and displeasing behaviors came from the Spouse Observation Checklist (SOC), a 400-item listing of relationship behaviors that spans 12 categories of dyadic functioning. Each evening, spouses read through the checklist and indicated which pleasing or displeasing events had occurred during the previous 24-hour period. A pair of trained observers used the Marital Interaction Coding System (MICS) to assess communication skillfulness from the videotaped negotiation sessions. The 29 MICS codes include verbal and nonverbal behaviors that have been a priori assigned to positive (e.g., approval, agree), negative (e.g., criticize, no response), or neutral (e.g., question) summary scores. All of these assessment procedures, with the exception of the Locke-Wallace, were developed at the University of Oregon and Oregon Research Institute and are described elsewhere (c.f., Weiss & Margolin, 1977).

Requests for reprints should be sent to Gayla Margolin, who is now at the Department of Psychology, University of Southern California, Los Angeles, California 90007.

Table 1
Correlations Among Measurements for Assessing Positive and Negative Dimensions of Marital Relationships

Measure	1	2	3	4	5	6
Self-report data						
1. Locke-Wallace adjustment scores	(.90)**					
2. Areas of Conflict scores	-.43*	(.83)**				
Spouse observer SOC data						
3. Pleasing behaviors	.40*	-.04	(.94)**			
4. Displeasing behaviors	-.23	.25	.13	(.78)**		
Trained observer MICS*						
5. Positive communication behaviors	.02	.31	.21	.04	(.70)**	
6. Negative communication behaviors	.01	-.11	-.13	-.06	-.77**	(.83)**

Note. Within-method correlations are in boldface; split-half reliability values corrected by the Spearman-Brown formula are in parentheses; and the remaining values are across-method correlations. Mean frequencies for odd and even days were correlated for Spouse Observation Checklist (SOC) reliability values. Mean frequencies for two 10-minute segments were correlated for Marital Interaction Coding System (MICS) reliability values. $n = 27$; scores are based on husband-wife averages.

* Interobserver agreement = 83.8%.

* $p \leq .05$, two-tailed.

** $p \leq .001$, two-tailed.

The correlational matrix in Table 1 contains split-half reliability coefficients, within-method correlations, and across-method correlations. Split-half reliability coefficients were consistently significant. Within-method correlations revealed significant correlations for two of the three methods investigated: Adjustment scores were inversely related to conflict scores, and positive communication behaviors were inversely related to negative communication behaviors. However, the only significant correlation across methods was between marital adjustment and mean daily frequency of pleasing behaviors.

The results of this study illustrate the relative independence among several commonly used procedures to assess marital relationships. The hypothesis that each of the two behavioral measurements would correlate with a more traditional measure of global marital satisfaction was confirmed for SOC pleasing behaviors but not for communication behaviors. Although previous studies have shown that communication skills discriminate between distressed and nondistressed couples, data from this study on distressed couples do not indicate a relationship between observers' coding of communication skills and couples' subjective satisfaction ratings. Poor communication may be the earmark of marital distress, but it appears from these data that the intensity of the communication deficit does not correspond systematically with the level of overall distress.

The lack of correspondence between SOC pleasing and displeasing behaviors lends support to the repeated finding that these two dimensions vary independently. However, the low correlation between behavioral frequencies on the SOC and on the MICS was contrary to an earlier finding by Patterson, Hops, and Weiss (1975) on a sample of 10 distressed couples. Patterson et al.'s rank-order correlations between positive behaviors in the laboratory setting and daily reports of reinforcing exchanges at home were .85 for wives and .61 for husbands.

The overall lack of correspondence among measures used in this study is not an indication of invalid measurement procedures but rather is an indication that marital adjustment is not a unitary dimension. If marital adjustment does imply a set of independent dimensions, a thorough assessment must take account of these different dimensions and provide a profile analysis of each couple's strengths and weaknesses. From the results of this study, it appears necessary to use different data collection procedures to assess the various aspects of marital functioning. In light of these findings and the limited information about the psychometric properties of particular instruments, further investigation is needed to choose wisely from the wide range of available marital assessment options.

References

- Birchler, G. R., & Webb, L. J. Discriminating interaction behaviors in happy and unhappy marriages. *Journal of Consulting and Clinical Psychology*, 1977, 45, 494-495.
- Patterson, G. R., Hops, H., & Weiss, R. L. Interpersonal skills training for couples in early stages of conflict. *Journal of Marriage and the Family*, 1975, 37, 295-302.
- Weiss, R. L., & Margolin, G. Marital conflict and accord. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook for behavioral assessment*. New York: Wiley, 1977.

Received July 26, 1977

Effects of Uncertainty About the Behavior of a Phobic Stimulus on Subjects' Fear Reactions

John Lick

State University of New York at Buffalo

Mark Condiotte

University of Oregon

Thomas Unger

State University of New York at Buffalo

Two groups of rat-phobic subjects were repeatedly exposed to a live rat on a nearby platform under instructional conditions designed to induce different expectancies about the probability of the rat remaining on the platform. Although the rat's behavior was the same in both conditions, subjects who were led to believe that the rat might leave the platform evidenced significantly greater cognitive and physiological arousal to repeated rat exposures than did subjects who were informed that the rat would definitely not leave the platform.

An important methodological issue in laboratory research dealing with "small animal phobia" concerns stimulus differences between the laboratory and natural environment that may affect subjects' uncertainty about how a phobic stimulus will behave. For example, in the laboratory, subjects are usually exposed to caged animals in a context designed to foster the expectation that events are "under control" and nothing unpredictable will happen. In contrast, outside the laboratory, subjects are likely to encounter uncaged animals moving in an unpredictable fashion. The study reported here is designed to investigate the reactions of phobic subjects toward a fear-eliciting animal as a function of instructions designed to affect uncertainty about the behavior of that animal.

Subjects were 19 females from introductory psychology classes who received research credit for their participation. They were selected from a pool of 600 females who completed a questionnaire designed to assess fear of rodents.

When subjects arrived at the laboratory, they were told that the experiment involved investigating their physiological reactions to various stimuli. Subjects were then connected to a polygraph.

After a baseline period of 10 min, the experimenter entered the subject's room and told her that she would be exposed to two different stimuli.

The first would be a roll of masking tape, and the second a live, harmless rat. Subjects were informed that the experimenter would place each stimulus on the platform, leave the room for 25 sec, and then return to remove the stimulus. At this point subjects were given an opportunity to withdraw from the experiment, and, if they decided to continue, they were asked to sign a consent form. Three minutes after signing this form, the experimenter exposed subjects to the masking tape and the live rat, with an interstimulus interval of 2 min between presentations. (The interstimulus interval for all subsequent rat exposures was 25 sec, and subjects were asked to rate their cognitive anxiety on a scale from 1 to 10 between trials.) Subjects who showed greater physiological reactivity to the rat than to the masking tape were retained in the study and were randomly assigned to one of two experimental conditions.

In the certainty condition ($n = 10$), subjects were told that the rat they would be repeatedly exposed to for the rest of the experiment had stayed on the platform for hundreds of hours and could not leave it. To enhance the credibility of this message, subjects were shown an electric grid on the platform and were told that this grid had been used to successfully train the animal to stay there.

In the uncertainty condition ($n = 9$), subjects were informed that the rat would "probably" not leave the platform; however, in the highly unlikely event that this did happen, subjects were told that the experimenter would intervene immediately to capture the animal.

During each of the last 10 rat exposures, the

Requests for reprints and for an extended report of this study should be sent to John Lick, Department of Psychology, State University of New York at Buffalo, 4230 Ridge Lea Road, Buffalo, New York 14226.

experimenter rated the rat's activity level on a 5-point scale through a small window. After the last exposure, subjects were disconnected from the polygraph and were given a behavioral avoidance test. Following this, subjects completed the rodent questionnaire again and an additional question designed to assess how confident they were during the experiment that the rat would remain on the platform.

Preliminary analyses indicated that (a) there were no significant between-groups differences on any measures before the introduction of the instructional manipulation defining the independent variable, (b) the instructional manipulation was highly effective in inducing appropriate cognitive expectancies vis à vis the rat's probable behavior, and (c) the animal's activity level was not confounded with either trial or experimental conditions.

Subjects' cognitive and physiological responses to the rat following the induction of differential expectancies were analyzed by repeated measures analyses of variance. The results of this analysis for the cognitive data indicated significant main effects for condition, $F(1, 17) = 21.1$, $p < .001$; trials, $F(9, 153) = 14.3$, $p < .001$; and the Condition \times Trials interaction, $F(9, 153) = 3.61$, $p < .001$. The results for the electrodermal response data show significant main effects for condition, $F(1, 17) = 17.1$, $p < .001$, and trials, $F(9, 153) = 7.43$, $p < .001$, but no significant Trials \times Condition interaction, $F(9, 153) = 1.19$, $p > .30$. Similarly, the heart rate data showed significant main effects for condition, $F(1, 17) = 7.85$, $p < .02$, and trials, $F(9, 153) = 6.82$, $p < .001$, and no significant Trials \times Condition interaction, $F(9, 153) = 1.02$, $p > .50$. Overall, these results show that subjects in the uncertainty condition manifested significantly more cognitive and physiological fear during the 10 rat exposures than subjects in the certainty condition. In addition, subjects in the certainty condition showed a significantly faster rate of cognitive fear habituation than uncertainty condition subjects.

Finally, results from the behavioral avoidance test indicated that subjects in the uncertainty condition completed an average of 9.9 steps (higher numbers indicating more approach), whereas certainty subjects completed a mean of 5.9 steps, $t(17) = 3.55$, $p < .002$. In contrast, pre-post differences on the rodent questionnaire failed to discriminate between the two experimental conditions, $t(17) = 1.45$, $p > .15$.

The results of this study appear to have several implications for laboratory research with subjects who are afraid of small animals. First, the results call into question the external valid-

ity of the laboratory-based behavioral avoidance test. Since this test assesses subjects' fear reactions to a caged phobic stimulus in an experimental context associated with safety and predictability, it is possible that subjects might show fear reductions on this measure without manifesting concomitant fear reductions to uncaged animals moving in unpredictable ways outside the laboratory. Although this issue requires additional study, the available information suggests that this lack of generalization may not be uncommon (e.g., Lick & Unger, 1975).

This study also has implications for research investigating habituation phenomena (e.g., Watson, Gaiend, & Marks, 1972) and paradoxical fear enhancement (e.g., Stone & Borkovec, 1975). Specifically, the significant Trials \times Condition interaction for the cognitive fear measure suggests that the slope of the curve describing subjects' habituation to repeated presentations of a phobic stimulus can be influenced by situational variables affecting the threat value of that stimulus. In this respect, cues that affect subjects' cognitive appraisal of the probability of a phobic stimulus suddenly approaching them may be particularly important in determining level of threat. This is important, since studies investigating habituation and fear-enhancement phenomena have used phobic stimuli consisting of either caged animals or slides, and subjects may habituate to these stimuli substantially faster than they would to stimuli that are much more threatening (e.g., encountering a moving animal in close proximity).

Finally, the significant effect of the instructional manipulation on subjects' behavioral avoidance test performance justifies additional attention because it suggests that the effectiveness of fear-reduction techniques using in vivo exposure might be enhanced by structuring some of these exposures under conditions more closely approximating those characterizing naturalistic encounters with phobic stimuli.

References

- Lick, J. R., & Unger, T. External validity of laboratory fear assessment: Implications from two case studies. *Journal of Consulting and Clinical Psychology*, 1975, 43, 864-866.
- Stone, N. M., & Borkovec, T. D. The paradoxical effect of brief CS exposure on analogue phobic subjects. *Behaviour Research and Therapy*, 1975, 13, 51-54.
- Watson, J. P., Gaiend, R., & Marks, I. M. Physiological habituation to continuous phobic stimulation. *Behaviour Research and Therapy*, 1972, 10, 269-278.

Received August 11, 1977

Some Useful Statistics for the Interpretation of the WISC-R

Terry B. Gutkin
University of Nebraska—Lincoln

A verbal comprehension deviation quotient (VCDQ) and a perceptual organization deviation quotient (PODQ) are presented as alternatives to the Verbal and Performance IQs when attempting to measure the Wechsler Intelligence Scale for Children-Revised Verbal Comprehension and Perceptual Organization factors. The superiority of the VCDQ and PODQ rests on their equally high reliabilities, equal ease of computation, and improved factorial validity when compared with the Verbal and Performance IQs. To aid in scatter analysis research, techniques that permit a determination of whether individual subtests deviate significantly from the subtest mean of the VCDQ and PODQ are discussed. Formulas are also presented for the computation of a freedom from distractibility deviation quotient (FDDQ) to measure the Freedom from Distractibility factor. Unlike the VCDQ and PODQ, an inconsistent pattern of factor loading mandates the use of age-specific formulae for the FDDQ. Reliability data indicate that the FDDQ should be used primarily for research purposes and only with considerable caution in applied settings.

Kaufman's (1975) factor-analytic investigation of the Wechsler Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974) standardization data has provided psychologists with a well-grounded empirical framework within which WISC-R scores can be analyzed with a maximum of objectivity. Among the study's major findings were the discovery of Verbal Comprehension, Perceptual Organization, and Freedom from Distractibility factors, all of which appeared across the entire age range of the normative sample. Kaufman advocates that the Verbal IQ be taken as a rough equivalent for the Verbal Comprehension factor. Unfortunately, the inclusion of the Arithmetic subtest score in the computation of the Verbal IQ, in conjunction with the failure of the Arithmetic subtest to show any substantive loadings on the Verbal Comprehension factor, results in the Verbal IQ being an impure measure of the Verbal Comprehension factor.

A better measure of this factor would be a linear combination of only the Information (I), Similarities (S), Vocabulary (V), and Comprehension (C) scores, as all of these subtests show substantial loadings on the Verbal Comprehension factor. Working from procedures outlined by Guilford (1954), it was determined

that such a linear combination would be as reliable as Wechsler's (1974) Verbal IQ. In fact, the reliabilities for the proposed linear combination and the Verbal IQ never differ by more than .02 of a point at any age level between 6½ and 16½. A simple formula for the conversion of this factor score into a verbal comprehension deviation quotient (VCDQ), with a mean of 100 and a standard deviation of 15, was developed with the aid of procedures discussed by Tellegen and Briggs (1967). $VCDQ = 1.47 (I + S + V + C) + 41.2$.

In a similar vein, it is suggested that the Performance IQ is not the best measure of a subject's perceptual organization skills. The Coding subtest, which is included in the computation of the Performance IQ, shows only trivial loadings on the Perceptual Organization factor at every age level. Using the Tellegen and Briggs (1967) procedures, the following formula was derived for computing a perceptual organization deviation quotient (PODQ) with a mean of 100 and a standard deviation of 15. $PODQ = 1.60 (PC + PA + BD + OA) + 36.0$. (PC = Picture Completion; PA = Picture Arrangement; BD = Block Design; OA = Object Assembly.) Reliability coefficients for the PODQ were computed for ages 6½-16½ and were never more than .01 of a point discrepant with those found by Wechsler (1974) for the Performance IQ.

With equivalent reliability, equal ease of computation, and superior factorial validity the VCDQ and PODQ should prove to be more

Requests for reprints should be sent to Terry B. Gutkin, Department of Educational Psychology, 130 Bancroft Hall, University of Nebraska, Lincoln, Nebraska 68588.

appropriate measures of verbal and nonverbal intellectual skills than the more traditional Verbal IQ and Performance IQ scores. The impact of using the VCDQ and PODQ versus the Verbal IQ and Performance IQ was examined using the WISC-R scores of all children who were referred for psychological examinations in a south-west urban school district over the period of a full school year ($N = 275$, average Full Scale IQ = 76, $SD = 13$). The results showed that despite very high correlations between the VCDQ and Verbal IQ ($r = .98$) and between the PODQ and Performance IQ ($r = .96$), important score shifts were quite common. Wechsler (1974) stated that verbal-performance discrepancies of 15 points or larger are "important" and call for further investigation. When comparing the magnitude of verbal-performance discrepancies using the VCDQ-PODQ scores instead of the Verbal IQ-Performance IQ scores, 8% of the students changed from either an important to an unimportant discrepancy or vice versa. An additional 16% of the group changed from either a nonsignificant to a significant discrepancy (a difference of 12 points or more) or vice versa. A further indication that the use of the VCDQ and PODQ scores in lieu of the Verbal IQ and Performance IQ scores may substantially affect test interpretation comes from the finding that 20% of the referred students had a score differential between either the VCDQ and Verbal IQ or the PODQ and Performance IQ of 7 points or more.

Since many of the WISC-R subtests have adequate subtest specificity (Kaufman, 1975), the investigation of subtest scatter may be of interest to researchers. One common technique of scatter analysis has been to determine if any particular subtest score is significantly different from the average of other subtest scores that purport to measure similar abilities (Davis, 1959).

Tables 1 and 2 provide the difference scores needed to determine if any of the individual

Table 1

*Minimum Deviations from the VCDQ
Subtest Average Required for Statistical
Significance*

p	I	S	V	C
.05	2.07	2.24	2.02	2.29
.01	2.72	2.94	2.66	3.02

Note. VCDQ = verbal comprehension deviation quotient; I = Information; S = Similarities; V = Vocabulary; C = Comprehension.

Table 2

*Minimum Deviations from the PODQ
Subtest Average Required for Statistical
Significance*

p	PC	PA	BD	OA
.05	2.48	2.62	2.18	2.77
.01	3.27	3.45	2.87	3.65

Note. PODQ = perceptual organization deviation quotient; PC = Picture Completion; PA = Picture Arrangement; BD = Block Design; OA = Object Assembly.

subtest scores that comprise the VCDQ and PODQ are significantly different from the average of all the subtest scores that comprise the VCDQ and PODQ. For example, if a subject's scores on the PODQ subtests are PC = 8, PA = 14, BD = 9, and OA = 9, the average for these subtests is equal to 10. The deviation from the PODQ average for each subtest is as follows: PC = -2, PA = 4, BD = -1, OA = -1. Table 2 shows that only the PA score differs significantly from the mean of the PODQ subtests ($p < .01$). The specific interpretation of any difference obtained in this manner will hinge on whether or not the subtest in question has adequate subtest-specific variance at the age level of the subject taking the test and on an analysis of what a particular subtest purportedly measures.

Although Sattler (1974) has presented a formula for the calculation of a freedom from distractibility deviation quotient (FDDQ), it is of questionable validity for several of the age levels that are encompassed by the WISC-R. Specifically, the Sattler formula is based on a linear composite score for the Arithmetic (A), Digit Span (DS), and Coding (Co) subtests. Even though these subtests do show substantial factor loadings on the Freedom from Distractibility factor at most age levels, the Co subtest has trivial loadings at the 6½-, 7½-, 14½-, and 16½-year levels (Kaufman, 1975). Sattler's formula, $FDDQ = 2.2 (A + DS + Co) + 34.0$, is thus appropriate for all age levels other than 6½, 7½, 14½, and 16½. The Tellegen and Briggs (1967) technique was used to compute an alternate formula for use at these four specific age levels: $FDDQ = 2.94 (A + DS) + 41.2$. It is important to note that for all age groups, the reliability of the FDDQ falls below .90 (ranging from a low of .83 at the 14½-year level to a high of .88 at the 12½-year level). Generally, .90 is accepted as the minimum reliability co-

efficient for scores that are used to discriminate among individuals (Nunnally, 1967). The FDDQ should thus be used primarily for research purposes and only with considerable caution in applied settings. The user must also keep in mind that unlike the VCDQ and the PODQ, the FDDQ measures a behavioral rather than a cognitive domain.

References

- Davis, F. B. Interpretation of differences among averages and individual test scores. *Journal of Educational Psychology*, 1959, 50, 162-170.
- Guilford, J. P. *Psychometric methods* (2nd ed.). New York: McGraw-Hill, 1954.
- Kaufman, A. S. Factor analysis of the WISC-R at 11 age levels between 6½ and 16½ years. *Journal of Consulting and Clinical Psychology*, 1975, 43, 135-147.
- Nunnally, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.
- Sattler, J. S. *Assessment of children's intelligence*. Philadelphia: Saunders, 1974.
- Tellegen, A., & Briggs, P. F. Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology*, 1967, 31, 499-506.
- Wechsler, D. *Manual for the Wechsler Intelligence Scale for Children-Revised*. New York: Psychological Corporation, 1974.

Received September 27, 1977 ■

Symptom Contamination of the Schedule of Recent Events

Robert E. Lehman
University of Idaho

Three hundred fifty-seven subjects completed the Schedule of Recent Events (SRE) and a widely used symptom scale. The correlation between unit weighted SRE scores and the symptom measure was significant and did not improve when the SRE scores were weighted by life change units. Four items from the SRE were identified as possibly representing symptoms in and of themselves. Correlations between each SRE item and the symptom measure were calculated, and it was found that these four items correlated more highly with the symptom scores than all but two of the remaining items from the SRE. It is suggested that because certain SRE items apparently represent symptoms, observed correlations between the SRE and various symptom measures may be artifactually high.

A growing body of research has demonstrated that an accumulation of life events may impair psychological functioning and may contribute to the development of psychopathology (Dohrenwend, 1973; Myers, Lindenthal, Pepper, & Ostrander, 1972; Paykel et al., 1969; Vinokur & Selzer, 1973). The Schedule of Recent Events (SRE), developed by Holmes and Rahe (1967), has been widely used to assess the accumulation of life events in such studies. The scale comprises 43 items that represent changes in personal behavior and life circumstances. Each item is assigned a weight termed the *life change unit* (LCU), based on the average of judges' estimates of the adjustive demand required by each event. These range from a low of 11 to a high of 100. The total score is the sum of the weights of those items checked as having occurred within the last 12 months. Numerous studies have demonstrated significant correlations between SRE scores and the development of both psychological and physical dysfunctions. (See review by Rahe, 1972.)

However, Rahe (1974) has noted that the LCU weightings, as compared to unit weightings, are not important in subject populations that report low to moderate magnitude events. One purpose of the present study was to further test this observation.

In addition, examination of the individual items on the SRE suggests that some of the

items may, themselves, represent symptoms of disturbed functioning. For purposes of the present study, four items were identified as generally symptomatic of a wide range of disorders. These are "change in eating habits," "change in sleeping habits," "change in social activities," and "revision of personal habits." To the extent that these items on the SRE also represent symptoms, correlations between SRE scores and symptom measures will be artifactually high. A second purpose of the present study was to further explore this possibility.

Three hundred fifty-seven undergraduate students at the University of Idaho were administered the SRE and a widely used symptom scale (Langner, 1962) that has been found to correlate with the SRE. The mean LCU of items checked by these subjects was 23.94 ($SD = 4.57$), which is quite close to the mean of approximately 25 LCUs for Rahe's (1974) navy subjects. Scores on Langner's symptom scale had a mean of 4.29 ($SD = 3.00$). Subjects were blind to the purpose of the experiment, and the order of administration of the scales was counter-balanced.

As anticipated for this low-LCU sample, the use of LCU weightings did not improve the correlations between the SRE and symptom scale. The product-moment correlations for weighted and unweighted scores were almost exactly the same ($r = .23$, and $r = .22$, respectively). Both were statistically significant ($p < .001$, two-tailed). These results further confirm previously observed relationships between SRE scores and disturbed functioning. They also support Rahe's (1974) contention that "when using the LCU magnitude scale with a sample of

Requests for reprints and for an extended report of this study should be sent to Robert E. Lehman, Department of Psychology, University of Idaho, Moscow, Idaho 83843.

subjects who report primarily low to moderate LCU events . . . one can dispense with the LCU scale" (p. 82).

A likely explanation for this phenomenon is that the LCU weights of the items most frequently affirmed by low-LCU-magnitude subjects do not differ a great deal. For example, the lower half of the scale (i.e., that half of the scale with the lowest LCU magnitude weights) represents less than one third of the range of the entire scale. Thus, because the LCU weights of these items are relatively similar, use of the weights may provide little advantage.

To assess the possibility that certain items on the SRE represent symptoms and may thus inflate correlations between SRE scores and measures of maladjustment, the correlations between each item from the SRE and the symptom scale were calculated. The resulting point-biserial correlations ranged from $-.063$ to $.202$. The four items mentioned above as probable symptoms were isolated and were found to have correlations ranging from $.133$ to $.202$, with a mean of $.167$. All were significant at the $.02$ (two-tailed) level. The mean correlation of the remaining items, in contrast, was only $.048$. Furthermore, the four "symptom" item correlations were larger than all but two of the remaining items. This was in spite of the fact that the symptom items had a considerably lower mean LCU score ($M = 18.25$) than the remaining items ($M = 36.34$). The fact that these four symptom items generally exhibit much higher correlations with the symptom scale than do the other items suggests that they strongly contribute to observed relationships between SRE scores and symptom measures.

To the extent that the SRE contains items that are symptomatic of disturbed functioning, correlations between the SRE and measures of symptomatology will be artifactually high. Indeed, Hudgens (1974) has noted that as many

as 29 of the 43 SRE items could represent symptoms or consequences of illness. It is not suggested that the observed detrimental effects of stressful life events on physical and psychological health are entirely artificial by-products of the existence of symptom items on the SRE. However, it is probable that they do contaminate the SRE. This possibility warrants attention in future research.

References

- Dohrenwend, B. Life events as stressors: A methodological inquiry. *Journal of Health and Social Behavior*, 1973, 14, 167-175.
- Holmes, T., & Rahe, R. The Social Readjustment Rating Scale. *Journal of Psychosomatic Research*, 1967, 11, 213-218.
- Hudgens, R. Personal catastrophe and depression. In B. S. Dohrenwend & B. P. Dohrenwend (Eds.), *Stressful life events: Their nature and effects*. New York: Wiley, 1974.
- Langner, T. A twenty-two item screening score of psychiatric symptoms indicating impairment. *Journal of Health and Human Behavior*, 1962, 3, 269-276.
- Myers, J., Lindenthal, J., Pepper, M., & Ostrander, D. Life events and mental status: A longitudinal study. *Journal of Health and Social Behavior*, 1972, 13, 398-406.
- Paykel, E., et al. Life events and depression. *Archives of General Psychiatry*, 1969, 21, 753-760.
- Rahe, R. Subjects' recent life changes and their near-future illness susceptibility. *Advances in Psychosomatic Medicine*, 1972, 8, 2-19.
- Rahe, R. The pathway between subjects' recent life changes and their near-future illness reports: Representative results and methodological issues. In B. S. Dohrenwend & B. P. Dohrenwend (Eds.), *Stressful life events: Their nature and effects*. New York: Wiley, 1974.
- Vinokur, A., & Selzer, M. Life events, stress and mental distress. *Proceedings of the 81st Annual Convention of the American Psychological Association*, 1973, 8, 329-330. (Summary)

Received October 13, 1977 ■

The Polydrug Assessment Scale: A Psychometric Technique for the Indirect Measurement of Drug Use

Khalil A. Khavari and Frazier M. Douglass IV

Midwest Institute on Drug Use, University of Wisconsin—Milwaukee

Presently, there are no psychometric instruments for the indirect measurement of polydrug use. To ameliorate this situation, self-report data from 335 college students were examined to identify a set of personality-attitudinal items that correlated best with polydrug use. Twenty items emerged as reasonably stable correlates of drug use. Additional analysis showed that the 20 items, when treated as a scale, were reliable, diagnostically accurate, and applicable to either sex. Thus, this work represents an initial step in the development of a scale for indirect measurement of an individual's extent of involvement with use of psychotropics.

Psychometric instruments are valuable tools in the area of substance abuse. Several instruments and procedures have been developed to assess alcoholism (Jacobson, 1976) and narcotic addiction (Siegel, 1976). However, no indirect instrument has been available for measurement of polydrug use. Thus, the purpose of the present study was to develop a psychometric instrument to measure indirectly the extent of each person's involvement with a variety of psychoactive substances.

Data for this purpose were obtained from 335 college undergraduates, who voluntarily completed a brief questionnaire and returned it anonymously. The questionnaire contained 132 statements (similar to Minnesota Multiphasic Personality Inventory items) presented with a Likert response format and 19 questions about current usage of various licit and illicit drugs. Responses from the drug use questionnaire were transformed to *T* scores and were averaged to yield a criterion measure of polydrug usage. (See Douglass & Khavari, in press.)

To select items for the Polydrug Assessment Scale (PAS), responses to the 132 items were

first correlated with the criterion. Then, all items with correlations less than .20 were excluded from further consideration. The remaining 50 items were then subjected to a series of multiple regression and cross-validation analyses until an accurate and stable set of 20 items was identified. Finally, these items were further examined to determine their reliability and their applicability to selected subgroups of respondents.

The resulting set of items are listed in Table 1. Their relationship with the criterion is reflected by bivariate correlations (absolute value) ranging from .21 to .51. When combined and regressed on the criterion in a random sample ($N = 166$), these items accounted for 61% of the variance associated with polydrug usage ($R = .78$, $p < .001$). Resulting regression coefficients were then used to cross-validate items using the remaining respondents. The sum of these weighted items (PAS score) accounted for 47% of the polydrug variance ($r = .69$, $p < .001$). Since results from these two analyses were comparable, all subjects were pooled for summary computations. The mean and standard deviation for the pooled group were 49.64 and 4.29, respectively. Multiple regression coefficients are presented in Table 1.

This study was supported by Research Grant DA01080 from the National Institute on Drug Abuse and funds from the College of Letters and Science and the Graduate School of the University of Wisconsin—Milwaukee to K. A. Khavari. We thank Phil Farber for his invaluable contributions to this work.

Requests for reprints and for an extended report of this study should be sent to K. A. Khavari, Department of Psychology, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin 53201.

Discriminative efficiency of the PAS was assessed by first dividing respondents into above average and below average polydrug users and then determining the correspondence between PAS scores and actual usage. A PAS score of 52, used as a dividing point, resulted in 81.5% correct classification; 24.18% were correctly identified as above average polydrug users,

57.31% were correctly identified as below average users, 6.57% were false positives, and 11.94% were false negatives. An increase or decrease in the PAS cutting point resulted in different percentages of correct and error classifications.

To examine the clinical utility of the PAS, respondents were divided into three groups (greater than +1 *SD*, between +1 and -1 *SD*, and less than -1 *SD*). Univariate analyses of variance showed that these three groups were significantly different in terms of all drugs except anti-infectious agents. People in the highest PAS group used every type of drug (except anti-infectious) more frequently than people with average PAS scores, and people with average scores used drugs more frequently than people with low PAS scores.

Examination of sex differences showed that males and females were not significantly different in terms of overall polydrug criterion scores or in terms of PAS scores. Thus, although males were more likely to use alcohol, marijuana, hashish, opiates, and cocaine, and females were more likely to use anti-infectious drugs, diet pills, tobacco, tranquilizers, sedatives, and relaxants, these differences are adequately controlled so that the PAS is applicable with either sex.

Since five of the PAS items specifically refer to drugs, some critics might contend that these items should not be included in the scale. Therefore, the multiple regression-cross-validation analyses were repeated after omitting the five items. The resulting correlation coefficients (.75 and .63, respectively) were not substantially different from values resulting from the full scale.

Reliability of the PAS was determined by applying the Spearman-Brown prophecy formula and by conducting a test-retest study with an additional group ($n = 17$). The Spearman-Brown reliability that was determined in the original group was .80, and the test-retest correlations were above .84 for most items (see Table 1).

These results show that the 20 items having low individual correlations with polydrug use can be combined to produce a stable instrument for indirectly assessing the degree of polydrug usage in males and females. Thus, the present scale is valuable for polydrug diagnostic purposes. It could be further developed by determining its use as a predictive device and identi-

Table 1

Polydrug Assessment Scale Items and Their Bivariate Correlation With Polydrug Use (P), Multiple Regression Weights (RC), and Test-Retest Coefficients (TR)

Item	P	RC	TR
1. I have participated in a political demonstration.	.37	.104	.85
2. I enjoy going to church.	-.34	-.27	.65
3. I have stolen (or shoplifted) something.	.38	.38	.86
4. My friends have been in trouble with the police.	.37	.22	.73
5. I have thought about suicide.	.26	.19	.90
6. I deserve severe punishment for my sins.	-.21	-.36	.99
7. I have a cough most of the time.	.27	.87	.85
8. I have few or no pains.	-.22	-.32	.87
9. Christ performed miracles such as changing water to wine.	-.26	.17	.76
10. I frequently notice my hand shakes when I try to do something.	.25	.49	.99
11. I have been quite independent and free from family rule.	.22	.26	.83
12. I have never been in trouble with the law.	-.41	-.56	.94
13. My parents have often objected to the kind of people I went around with.	.37	.42	.85
14. I played hooky from school quite often as a youngster.	.31	-.26	.87
15. Religion is the most important system of human values.	-.29	-.19	.83
16. Most people who use marijuana lead a normal life.	.36	.36	.79
17. People should only use drugs for medical reasons.	-.51	-.68	.84
18. The facts on crime and drug abuse show we will have to crack down harder on criminals and drug addicts.	-.35	-.56	.85
19. People have the right to the pursuit of happiness and well being. If to achieve this requires taking illicit drugs, that's their business.	.37	-.55	.86
20. Drugs are there to be used and enjoyed by people.	.46	.59	.94

Note. Decimals have been omitted. Multiple regression constant = 50.12.

fyng new items that substantially improve its assessment capability.

References

Douglass, F. M., & Khavari, K. A. The drug use index: A measure of the extent of polydrug usage. *International Journal of the Addictions*, in press.

Jacobson, G. R. *The alcoholisms: Detection, diagnosis and assessment*. New York: Human Sciences Press, 1976.

Siegel, J. Psychological testing and the detection of drug abusers. *International Journal of the Addictions*, 1976, 11, 1031-1043.

Received October 18, 1977 ■

Comparison of Symbolic and Overt Aversion in the Self-Control of Smoking

Oscar A. Barbarin
University of Maryland

This research compares rapid smoking (overt aversion), covert sensitization (symbolic aversion), and a combination of the two in a self-punishment procedure for reducing and eliminating cigarette smoking. Sixty adults (M age = 40) were randomly assigned to one of three experimental groups or to the control group. Ten training sessions were spaced over a 1-month period. The experimental groups smoked significantly less than the control group at each follow-up point. The overt group achieved significantly greater reduction than the symbolic group, and the combined group did not differ significantly from either of the single treatment groups. One year after treatment, 6 of 15 persons in the overt aversion group were completely abstinent, as opposed to 1 in each of the other experimental groups.

The present study examines the relative efficacy of symbolic and overt aversion in self-control procedures designed to eliminate smoking. Sixty adults (M age = 40) who smoked at least one pack of cigarettes a day were medically screened and assigned to one of three experimental groups (rapid smoking, symbolic aversion, combined treatment) or a control group. A week of self-monitoring prior to treatment provided the baseline against which treatment effects were to be measured. Training in each self-control strategy was conducted in 10 1-hour sessions spaced over a 1-month period. Treatment was carried out in groups of 3-7 persons. At each session participants were asked to report the number of cigarettes smoked since the previous session and whether they had used the self-control procedure to avoid normal smoking.

After 10 minutes of relaxation exercises, the overt aversion group began rapid smoking, which called for normal inhalations every 6 sec. Participants were paced by a prerecorded tone and a voice signal that simultaneously counted out consecutive numbers every 6 sec. Trial length and the latencies between trials were determined

by the participants. They were instructed to smoke as long as they could tolerate smoking. Group members recorded the number that they heard at the beginning and end of each trial, making it possible to calculate the number of inhalations per trial and the latencies between trials. At the end of each trial, participants also rated the aversiveness of their thoughts and physical reactions.

The symbolic aversion group was first asked to imagine as clearly as possible several scenes in which unpleasant consequences occurred as a result of smoking. Most scenes were drawn from situations listed by participants in their smoking logs compiled prior to treatment. Participants called to mind situations in familiar places such as homes, work, or with friends in which they frequently experienced the temptation to smoke. Then they were asked to imagine the occurrence of a variety of aversive consequences. In many instances these imagined consequences were isomorphic with the real consequences for the overt group, namely, nausea, coughing, dizziness, and discomfort in the chest. Other graphic images such as vomiting on self and uncontrollable coughing were also used to enhance the aversive nature of the imagined scenes. After the experimenter described an aversive scene, the participants imagined the scene. Following the aversive scenes, participants imagined an aversion relief scene that included the refusal to smoke. In these scenes participants imagined feeling calm, relaxed, and generally good about themselves. After each trial, members of the symbolic aversion group rated the vividness and

This study is based in part on a doctoral dissertation completed at Rutgers University, under the direction of Peter Nathan. Data analysis was made possible through a grant from the Computer Science Center at the University of Maryland.

Requests for reprints and for an extended report of this study should be sent to Oscar A. Barbarin, Community Field Station, Department of Psychology, University of Maryland, College Park, Maryland 20742.

Table 1

Mean Percentage of Baseline Smoking for Experimental Groups During Treatment and Follow-Up

Time period	Group								F
	Symbolic		Overt		Combined		Control		
	M	SD	M	SD	M	SD	M	SD	
	Treatment								
1 week	22	11	9	9	11	16	91	.12	58.41**
2 weeks	13	13	4	7	9	.09	88	20	86.36**
3 weeks	16	21	1	2	11	20	87	29	59.12**
Follow-up									
1 week	29	29	1	7	20	28	87	24	35.40**
2 weeks	42	36	8	13	36	48	90	22	19.27**
4 weeks	47	39	18	24	36	46	85	29	11.17**
8 weeks	55	39	21	17	50	48	90	23	9.37**
12 weeks	59	33	28	31	53	37	92	26	10.32**
1 year	84	28	56	40	75	39	100	0	4.38*

* $n = 14$.

^b $n = 15$.

* $p \leq .01$.

** $p \leq .001$.

aversiveness of the scenes. Participants were asked to use the symbolic aversion procedure and the relaxation exercises when they felt an urge to smoke. In addition, they were asked to practice for 10 minutes twice a day.

The combined aversion group used both symbolic and overt aversion, as described above. Participants used relaxation exercises, imagined unpleasant consequences for smoking, and engaged in rapid smoking while concentrating on these unpleasant images. Members of the control group were told that they could not be accommodated in the training groups but that they would be sent descriptions of self-control procedures used in the program and that their

progress would be monitored through telephone contacts. Subsequently, they were sent publications of the American Cancer Society describing a self-directed plan for monitoring and gradual reduction of smoking. The control group was contacted weekly for 1 month and at each of the follow-up points. During the 4 weeks of treatment, they were asked about their progress and were encouraged to maintain their efforts to abstain from smoking. All participants were contacted 1 week, 2 weeks, 1 month, 2 months, 3 months, and 1 year following treatment.

One participant each from the symbolic and combined groups dropped out before follow-up could be completed. The resulting groups did not differ significantly with respect to sex, age, cigarette consumption, or previous attempts to quit smoking. A one-way analysis of variance was performed on the percentage of baseline smoking reported at each follow-up point. Table 1 presents the means, standard deviations, and F values for smoking rates from treatment to the 1-year follow-up. Significant differences were found at each point between the experimental groups and the control groups. In addition, overt aversion produced greater reductions in smoking than symbolic aversion. However, there were no significant differences between the combined group and either the symbolic or overt aversion

Table 2

Number of Abstinent Participants During Follow-Up

Group	Weeks after treatment					
	1	2	4	8	12	152
Symbolic ^a	4	4	3	3	1	1
Overt ^b	14	12	12	8	6	6
Combined ^a	7	6	6	5	3	1
Control ^b	1	1	1	1	1	0

^a $n = 14$.

^b $n = 15$.

groups. Interestingly, this pattern held up even at the 1-year follow-up.

A similar pattern was observed with respect to the number of participants maintaining abstinence (see Table 2).

The overt aversion group appeared decidedly superior to all other groups with respect to long-term maintenance of abstinence.

On a 9-point scale, the symbolic aversion group rated their images as significantly clearer than did the overt aversion group ($M = 7.72$ vs. 6.00). However, the symbolic group did not differ from the combined group ($M = 6.67$). In rating the aversiveness of physical reactions during aversion training, the symbolic aversion group ($M = 6.8$) was significantly lower than either the overt ($M = 8.2$) or the combined groups ($M = 8.2$). The combined groups had less training with the rapid-smoking technique than the overt group with respect to the number of trials (25.4 vs. 31.4), the number of cigarettes per session (5.1 vs. 7.1) and the total time in rapid-smoking training (167 vs. 200 minutes). At the end of treatment, participants rated the usefulness of various treatment components. The overt and combined groups rated rapid smoking as most useful, with group support second. The symbolic group gave highest ratings to relaxation training and group support.

The results of the study point to the superiority of overt aversion over symbolic aver-

sion in the self-control of cigarette consumption. This is true both with respect to the absolute level of smoking reduction and the number of individuals who were able to maintain abstinence. However, each of the experimental treatment groups smoked significantly less than the control group. These findings suggest that symbolic aversion shows promise as a self-modification procedure for gaining control over smoking behavior. The failure of the combined treatment group to achieve a level of smoking reduction superior to that of the symbolic treatment alone is perplexing. However, similar findings are reported by Danaher (1977), who combined several treatment components. It is likely that participants attempting to apply several techniques may not be able to master each of them fully. In this case, differences in the amount of training in the overt condition may also contribute to the diminished effect of the combined treatment group. Consequently, additional attention must be given to the effects of combining treatment.

Reference

- Danaher, B. G. Rapid smoking and self-control in the modification of smoking behavior. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1068-1075.

Received October 19, 1977 ■

Moderators of Racial Differences on the MMPI

Arthur I. Rosenblatt and David A. Pritchard
University of Mississippi

Multiple discriminant analysis of Minnesota Multiphasic Personality Inventory (MMPI) scores between high-IQ white, high-IQ black, low-IQ white, and low-IQ black subjects yielded two significant canonical variates. The results suggest that racial differences on the MMPI do not occur in all racial comparisons but instead are restricted to low-IQ groups.

In his review of racial differences on the Minnesota Multiphasic Personality Inventory (MMPI), Gynther (1972) concluded that

the degree of MMPI differences between blacks and whites appears to be affected by such variables as education, residence and cultural separation. (p. 390)

The largest racial differences reportedly occur among poorly educated, rural, and isolated subjects, whereas such differences are attenuated or even eliminated among better educated, urban, and integrated subjects. The empirical basis for this conclusion is unclear, however, since only one of the studies reviewed by Gynther (Erdberg, 1970) contained a factorial design to test the moderating effect of one of the hypothesized variables (i.e., residence: urban vs. rural). The remaining studies generally used fixed values for education (e.g., ninth graders), for residence (e.g., urban), and for cultural separation (e.g., segregated school), and thus provided no test of the moderating effect of these variables.

Since Gynther's (1972) review, only one study has specifically investigated the effect of one of Gynther's hypothesized moderator variables. Davis and Jones (1974) performed analyses of variance on each of the MMPI clinical scales using race, diagnosis (schizophrenic vs. "other disorders"), and education (12+ grades vs. 11 or fewer grades completed) as factors. They reported no significant main effects for race on the clinical scales, but they did find a significant Race \times Education interaction on the

Pa and *Sc* scales, with the less educated blacks scoring higher than the other groups.

In a related analysis of the same data, Cowan, Watkins, and Davis (1975) reported that only the poorly educated black nonschizophrenics were significantly misclassified as schizophrenic when an empirical rule was used to classify MMPI profiles into schizophrenic and nonschizophrenic groups. Both of these analyses support the hypothesis that racial differences on the MMPI are moderated by educational status. The present study attempts to replicate this finding on a new sample, using a more appropriate method of statistical analysis.

The subjects were 104 black and 191 white male inmates at the Mississippi State Penitentiary, who were each administered the Wechsler Adult Intelligence Scale (WAIS) and the MMPI as part of an evaluation for vocational rehabilitation services. Subjects ranged in age from 17 to 59 (white: $M = 25.8$, $SD = 7.54$; black: $M = 24.5$, $SD = 5.83$), and in education from 3 completed grades to 16 completed grades (white: $M = 10.7$, $SD = 2.32$; black: $M = 11.2$, $SD = 2.02$). All subjects scored above the standard score of 70 on the reading portion of the Wide Range Achievement Test (WRAT) (white: $M = 101.2$, $SD = 15.81$; black: $M = 91.1$, $SD = 13.0$). Over 90% of the subjects were new arrivals at the prison at the time of testing, whereas the remaining subjects had reached their time of parole eligibility.

Highest grade completed was rejected as a potential moderator variable, since a majority of the subjects had been in school prior to integration in Mississippi, and thus highest grade completed might not indicate an equal educational experience for the two racial groups. Instead, the Full Scale WAIS IQ was adopted as an approximate indicator of the kinds of differences that might moderate racial differences on the

This article is based on a master's thesis completed at the University of Mississippi by the first author.

Requests for reprints should be sent to David Pritchard, Department of Psychology, University of Mississippi, University, Mississippi 38677.

MMPI. The product-moment correlations between WAIS Full Scale IQ and WRAT scores were .64 (Reading, $p < .001$), .54 (Spelling, $p < .001$), and .70 (Arithmetic, $p < .001$), whereas the correlations between highest grade completed and WRAT scores were .32 (Reading, $p < .001$), .40 (Spelling, $p < .001$), and .35 (Arithmetic, $p < .001$). Thus, Full Scale IQ appeared to be a better indicator of educational achievement than highest grade completed.

The sample was divided into four subgroups defined jointly by race and by IQ score (dichotomized at the sample median of 93.0). The four groups were labeled high-IQ white ($M = 106.0$, $SD = 8.81$, $n = 121$), high-IQ black ($M = 100.8$, $SD = 6.68$, $n = 23$), low-IQ white ($M = 85.7$, $SD = 5.69$, $n = 70$), and low-IQ black ($M = 81.2$, $SD = 6.47$, $n = 81$). A multiple discriminant analysis (Klecka, 1975) was performed on these four groups using K -corrected raw scores on the MMPI's 10 clinical scales as variables. If IQ functions as a moderator of racial differences on the MMPI, then the resultant canonical variates should separate the low-IQ white and low-IQ black groups but not the high-IQ white and high-IQ black groups. On the other hand, if IQ does not moderate racial differences on the MMPI, then canonical variates should separate the racial groups within both the high and low IQ categories.

The first variate produced a Wilk's lambda of .6958, which is approximately equal to a chi-square of 104.1 ($df = 30$, $p < .001$; Overall & Klett, 1972). The second variate produced a Wilk's lambda of .8508 ($df = 18$, $p < .001$), and the third variate yielded a nonsignificant Wilk's lambda of .9766 ($p < .560$). The first variate accounted for 18.23% of the MMPI variance and separated the high and low IQ groups; the second variate accounted for 12.89% of the variance and separated the low-IQ white group from the low-IQ black group. Accordingly, the first variate can be interpreted as an IQ effect on MMPI scores, and the second can be interpreted as a (moderated) racial effect on MMPI scores. In support of this interpretation, it should be noted that the greatest contributions to the second (moderated racial) variate were made by Scales 1, 3, 8, and 9, as indicated by the relative magnitudes of the standardized discriminant weights for these scales. Blacks scored higher than whites on Variate 2 for Scales 1, 8 and 9, whereas whites scored higher than blacks on Scale 3. This is precisely the

pattern of racial differences on the MMPI that previous studies have reported most often (Gynther, 1972; Rosenblatt, 1976). It can thus be concluded that IQ moderated the effect of race on MMPI scores and that such racial differences were almost exclusively limited to lower IQ subjects.

Since the above results may have been influenced by the inclusion of "invalid" profiles (Costello, Tiffany, & Grier, 1972), the analysis was repeated five times on successive samples of the 295 profiles. The samples differed from each other only in the maximum F score allowable for inclusion, and thus represented successively more stringent criteria for profile validity. Results from these five analyses confirmed that the results obtained in the analysis of all profiles were not due to the inclusion of invalid profiles. In fact, the exclusion of profiles with high F scores increased the size of the IQ effect and the moderated racial effect on MMPI clinical scales.

References

- Costello, R., Tiffany, D., & Grier, R. Methodological issues and racial (black-white) comparisons on the MMPI. *Journal of Consulting and Clinical Psychology*, 1972, 38, 161-168.
- Cowan, M., Watkins, B., & Davis, W. Level of education, diagnosis and race-related differences in MMPI performance. *Journal of Clinical Psychology*, 1975, 31, 442-444.
- Davis, W., & Jones, M. Negro versus Caucasian psychological test performance revisited. *Journal of Consulting and Clinical Psychology*, 1974, 42, 675-679.
- Erdberg, S. MMPI differences associated with sex, race and residence in a southern sample. (Doctoral dissertation, University of Alabama, 1969). *Dissertation Abstracts International*, 1970, 30, 5236B. (University Microfilms No. 70-9343)
- Gynther, M. White norms and black MMPIs: A prescription for discrimination? *Psychological Bulletin*, 1972, 78, 386-402.
- Klecka, W. Discriminant analysis. In N. Nie, C. Hull, J. Jenkins, K. Steinbrenner, & D. Bent (Eds.), *Statistical package for the social sciences* (2nd ed.) New York: McGraw-Hill, 1975.
- Overall, J., & Klett, C. *Applied multivariate analysis*. New York: McGraw-Hill, 1972.
- Rosenblatt, A. *A multivariate analysis of racial differences on the MMPI*. Unpublished master's thesis, University of Mississippi, 1976.

Received October 19, 1977 ■

Problems Associated With the Typological Measurement of Sex Roles and Androgyny

Jeffrey A. Kelly

Department of Psychiatry and Human Behavior
University of Mississippi Medical Center, Jackson

Wyndol Furman

University of Denver

Veronica Young

Belhaven College

Four new measures of sex role style (the Bem Sex-Role Inventory (BSRI) the Personal Attributes Questionnaire (PAQ) the Personality Research Form (PRF) ANDRO scale, and the Masculinity and Femininity scales of the Adjective Check List; ACL) each define sex role style and androgyny in similar conceptual and psychometric terms. Although these scales have often been used interchangeably, the current study is the first to examine interscale comparability among these inventories. Although correlations among the respective Masculinity and Femininity raw (continuous) scale scores on the BSRI, PAQ, PRF ANDRO scale, and the ACL were moderately high, a large proportion of subjects were classified into different sex role categories (masculine typed, feminine typed, androgynous, or undifferentiated), with the category depending on the inventory used. In fact, when corrected for chance agreements, the majority of subjects (61%) were actually categorized discrepantly by any pair of inventories. This suggests limited comparability of sex role research findings based on different inventories, and when sex role styles are dichotomized into broad typological quadrants, as is the current practice in sex role research, substantial predictive utility is lost.

Just as bipolar masculinity-femininity scales had once flourished, new sex role inventories based on a conceptual model using independent measurement of masculinity-femininity have proliferated recently. The Bem Sex-Role Inventory (BSRI; Bem, 1974), the Personal Attributes Questionnaire (PAQ; Spence, Helmreich, & Stapp, 1975), the PRF ANDRO scale (Berzins, Welling, & Wetter, 1978), and the Masculinity-Femininity scales of the Adjective Check List (ACL; Heilbrun, 1976) each define sex roles in similar terms, and each has been used to designate sex-typed androgynous, and undifferentiated roles. Because these four scales rely on highly similar definitions of sex roles, are all capable of designating androgyny, and all use similar scoring procedures to yield four sex role categories, the inventories have been

used almost interchangeably in recent sex role research.

However, Kelly and Worell (1977) have noted that while purporting to measure the same sex role constructs, each of these scales samples somewhat different content domains, were developed using different psychometric and item selection procedures, and were subjected to different criteria of validity and reliability. Further, although the four scales have been used interchangeably, no studies have directly compared these inventories. An important question is whether individuals categorized as sex typed, androgynous, or undifferentiated on one scale receive the same designation on another. If this were not the case, it would impose serious restrictions on the comparability of research findings based on different scales and would require further examination of configural sex role scoring procedures. The current study addresses the questions of interscale comparability and strategies of sex role categorization.

One hundred thirty (65 male and 65 female) undergraduate college students served as subjects in the study. All inventories were administered in fully counterbalanced order. Directions for completing each of the scales were those

This study was partially supported by National Institute of Mental Health Grant 1 RO1 MH29884-01A1 to the first author.

Requests for reprints and for an extended report of this study should be sent to Jeffrey A. Kelly, Department of Psychiatry and Human Behavior, University of Mississippi Medical Center, 2500 North State Street, Jackson, Mississippi 39216.

used by each respective inventory's author. However, because time limitations precluded the administration of the entire ACL, only those ACL adjectives actually comprising the Masculinity-Femininity scales were administered. These adjectives were listed in random order on a checklist form.

The scales were then scored to yield separate Masculinity and Femininity scores for each student on the BSRI, PAQ, ANDRO scale, and ACL. Following the procedure advocated by each scale's author, subjects were considered masculine typed if they scored above the Masculinity scale median and below the Femininity scale median on an inventory. Those subjects whose Femininity scores were above the median and whose Masculinity scores were below the median were categorized as feminine typed. Persons were considered androgynous if they exceeded an inventory's masculinity and femininity medians, and they were considered undifferentiated if they fell below both medians on an inventory.

Pearson product-moment correlations were first calculated among the raw masculinity and femininity scores of all inventories. For sexes combined, interscale masculinity score correlations were BSRI/PAQ = .85; BSRI/ANDRO = .70; BSRI/ACL = .75; PAQ/ANDRO = .66; PAQ/ACL = .70; and ANDRO/ACL = .61. Interscale femininity score correlations were BSRI/PAQ = .73; BSRI/ANDRO = .62; BSRI/ACL = .68; PAQ/ANDRO = .59; PAQ/ACL = .51; and ANDRO/ACL = .57. The mean correlation between the four inventories' raw Masculinity scales was .71, and the mean correlation between Femininity scales was .62.

Since current approaches to sex role measurement assign subjects to one of four typologies, interscale comparability was further assessed by determining the percentage of subjects who were classified into the same sex category by each pair of scales. A crucial question is whether the same individuals are assigned to the same sex role categories across inventories. Table 1 reveals that the percentage of agreement in individual subject classification was quite low, averaging only 56% of the subjects between any two scales' categories (range = 52% agreement to 61% agreement for sexes combined and 49% to 63% for sexes separately). Thus, an average of 46% of all subjects were assigned to discrepant sex role categories when the classification outcomes of two inventories were compared.

To correct these results for interscale classification agreements due to chance, Kappa co-

Table 1
Interscale Classification Agreement Rates

Inventories compared	% assigned to same category	% corrected for chance
BSRI/PAQ		
Sexes combined	60.8	37.9
Males	60.0	43.1
Females	61.5	45.9
BSRI/ANDRO		
Sexes combined	55.4	40.0
Males	52.3	28.1
Females	58.5	39.1
BSRI/ACL		
Sexes combined	56.2	42.0
Males	56.9	39.6
Females	55.4	35.7
PAQ/ANDRO		
Sexes combined	51.5	35.9
Males	49.2	28.9
Females	53.8	35.6
PAQ/ACL		
Sexes combined	55.4	40.0
Males	50.8	29.7
Females	60.0	44.0
ANDRO/ACL		
Sexes combined	56.2	42.2
Males	49.2	28.6
Females	63.1	47.4

Note. BSRI = Bem Sex-Role Inventory; PAQ = Personal Attributes Questionnaire; ANDRO = Personality Research Form ANDRO scale; and ACL = Adjective Check List.

efficients (Cohen, 1960) were determined. When the percentage of agreement between the classification outcomes of inventories have been adjusted for chance, the mean Kappa percentage agreement drops to 39. Thus, the majority of subjects are actually classified discrepantly when a second sex role inventory is used. When the percentage of agreement in subject classification across all four sex role inventories was calculated, only 30% of the subjects were found to be categorized the same on all four inventories.

It is apparent that persons classified as androgynous, sex typed, or undifferentiated using one sex role scale may well be a very different subsample than persons with the same designation based on another scale. This is the case even when the scales purport to assess similar characteristics, use the same scoring procedures, and assign subjects with identical labels based

on their scores. These findings raise serious limitations concerning the comparability of research results when differing inventories are used.

In the current study, relatively high correlations were obtained between the Masculinity and, to a lesser degree, the Femininity scales of the four inventories when subjects' scores were treated as continuous variables. However, when subjects' scores were dichotomized by median split categorizations, many, and in some cases, most subjects are discrepantly classified. Lost variance and reduced predictive utility are exacerbated by the prevalent scoring procedure in which subjects are classified into four sex role categories on both the Masculinity and Femininity scales of a given inventory. Although masculinity and femininity as defined by these scales appear to be unrelated and should be measured separately, the current data indicate that there is little empirical basis for configurally combining these scores into broad, typological categories based on median splits. Multiple linear measurement or multiple regression analyses may enable future research to more precisely examine the adjustive and social-behavioral implications of sex role orientations.

References

- Bem, S. L. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1974, 42, 155-162.
- Berzins, J. I., Welling, M. A., & Wetter, R. E. A new measure of psychological androgyny based on the Personality Research Form. *Journal of Consulting and Clinical Psychology*, 1978, 46, 126-138.
- Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, 20, 37-46.
- Heilbrun, A. B. Measurement of masculine and feminine sex role identities as independent dimensions. *Journal of Consulting and Clinical Psychology*, 1976, 44, 183-190.
- Kelly, J. A., & Worell, J. New formulations of sex roles and androgyny: A critical review. *Journal of Consulting and Clinical Psychology*, 1977, 45, 1101-1115.
- Spence, J. T., Helmreich, R., & Stapp, J. Ratings of self and peers on sex role attributes and their relationship to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 1975, 32, 29-39.

Received October 20, 1977 ■

Male and Female Treatment Differences: Can They Be Generalized?

Andrew C. Del Gaudio, Paul J. Carpenter, and Gary R. Morrow
Department of Psychiatry
University of Rochester Medical Center

The present study evaluated the generalizability of previous findings that male and female patients may receive differential psychiatric treatment. A comparison of 156 female and 66 male outpatients on demographic, clinical, and self-report measures of mood, symptoms, and interpersonal concerns revealed no sex differences. The treatment variables of length of therapy and the prescription of medication were the dependent measures. The results indicated that with a broad sample of male and female patients, there were no significant sex-related differences on the dependent variables. Thus, it is suggested that the generalization of sex-specific treatment differences be qualified.

A recent study by Stein, Del Gaudio, and Ansley (1976) found that a sample of neurotically depressed female outpatients had significantly more therapy sessions and were significantly more likely to be prescribed psychotropic medications, especially the more potent varieties, than were a sample of male neurotic depressives. These findings were obtained despite the fact that the male and female patients were indistinguishable on demographic, clinical, and self-report measures of symptoms, mood, and interpersonal concerns. Their data supported the findings of others (Abramowitz, Abramowitz, Roback, Corney, & McKee, 1976) that differential attitudes of clinicians, rather than patient factors, are a salient element in the psychotherapeutic process.

The present study examined the generalizability of the earlier study (Stein et al., 1976) by attempting to replicate the sex-specific findings with a different sample of patients representing a broader range of clinical diagnoses rather than a single category.

Subjects were 222 patients who were admitted to the same outpatient clinic used in the Stein et al. (1976) study. Demographic and clinical data were gathered on each patient, as well as their responses to the following measures of mood, symptoms, and interpersonal concerns: Terminator-Remainer Scale (TR), Hopkins

Symptom Rating Scale (HSRS, five subscales), Profile of Mood States (POMS, six subscales), FIRO-B (six subscales), and the Marlowe-Crowne Social Desirability Scale. The majority were white (84.6%), married (41.6%), middle class (68.3%), and female (70.3%).

A broad spectrum of psychiatric diagnoses was present in the sample: neurosis (37.9%); transient situational disorder (29.4%); personality disorder (15.8%); psychosis (7.7%); addictive states (2.3%); and deferred, unknown, and other diagnoses (6.9%); the other diagnoses chiefly included psychophysiologic disorders, organic brain syndrome, and mental retardation).

Chi-square comparisons of the male and female patients on the demographic variables of age, race, marital status, and education yielded no differences. Male-female differences in self-ratings of mood, symptoms, and interpersonal concerns examined by *t* tests revealed that (a) Female patients had significantly higher scores than males on the Depression subscale of the HSRS ($p < .05$); (b) males, on the other hand, had significantly higher depression scores than females on the POMS ($p = .05$); (c) males had significantly higher expressed control scores ($p < .01$) than did females on the FIRO-B.

Significantly elevated Depression scores for both male and female patients on different measures make it safe to assume that there are essentially no sex-specific differences on depression. The findings on the FIRO-B can be attributed to chance, since 1 significant result out of a possible 19 is within the statistical expectation of chance. Therefore, this study yielded no evidence that male and female psychiatric patients

Requests for reprints and for an extended report of this study should be sent to Paul J. Carpenter, who is now at the University of Southern Mississippi, P.O. Box 8482, Southern Station, Hattiesburg, Mississippi 39401.

entered treatment with significantly different levels or types of subjective distress and interpersonal needs and concerns, a result in agreement with the Stein et al. (1976) study.

The present investigation did not, however, reveal any sex differences in the average number of therapy sessions or the extent to which medications were prescribed. Thus, when the patient sample was expanded beyond a single diagnostic category (neurotic depression), no treatment differences emerged between male and female patients drawn from the same clinic population, as they did in Stein et al. (1976). This strongly suggests that limits be placed on the generaliza-

tion that differential psychiatric treatment is afforded to men and women.

References

- Abramowitz, S. I., Abramowitz, C. V., Roback, H. B., Corney, R. T., & McKee, E. Sex-role related countertransference in psychotherapy. *Archives of General Psychiatry*, 1976, 33, 71-73.
- Stein, L. S., Del Gaudio, A. C., & Ansley, M. Y. A comparison of female and male neurotic depressives. *Journal of Clinical Psychology*, 1976, 32, 19-21.

Received November 7, 1977 ■

Effect of Psychotomimetics (LSD and Dextroamphetamine) on the Use of Figurative Language During Psychoanalysis

Michael Natale
Department of Psychiatry
Columbia University

Michael Kowitt
Philadelphia Psychiatric Center
Philadelphia, Pennsylvania

Charles C. Dahlberg and Joseph Jaffe
Department of Communication Sciences
New York State Psychiatric Institute
New York, New York

This study examined the effect of LSD, dextroamphetamine, and a placebo on patients' use of figurative language in psychoanalytic sessions. A longitudinal design was performed on three male psychoanalytic patients who volunteered to receive each drug (LSD-50-100 u, dextroamphetamine-15 mg, placebo) seven times over a period of 1½ years. The patients, analyst, and raters were all blind concerning the drug condition. Forty-minute verbatim transcripts were scored for figurative language by two trained raters. Results showed that for two of the three patients, the use of nonliteral language, in particular, "novel" figurative phrases, was significantly increased by LSD.

The rationale for the clinical use of psychotomimetics (LSD, anticholinergics) is that they augment traditional psychotherapy by intensifying and accelerating transference, recall, loosening of defenses, and so forth (Abramson, 1967). Is there any empirical verification for these notions or do they remain presuppositions? Fink (1974) has presented evidence suggesting that psychotomimetics induce a decrease in syntactic "defensive language behaviors" (use of third-person reference and qualifiers, use of past tense instead of present tense). LSD-induced language changes are hypothesized to resemble the speech of acute schizophrenics. However, Gottshalk and Gleser (1969) presented contrary evidence to the effect that psychotomimetics do not promote "schizophrenic" speech, but they do promote

language that reflects a cognitive impairment (organicity).

In light of the above-cited contradictory findings, it is fair to state that the effects of LSD on the speech of normals is undetermined at present. The goal of this study was to clarify the effect of psychotomimetics on speech in (a) the examination of the *therapeutic* language effects of LSD, dextroamphetamine, and a placebo on patients' speech during actual psychotherapy sessions and (b) the use of a language measure that is significantly related to "insight" on the part of the patient. We (the experimenters) chose "figurative language" as a speech index of clinical change because investigators have become increasingly aware of the role of figurative language in psychotherapy; specifically, *novel* nonliteral language is associated with insight. (See Billow, 1977, for a review.) To sum up, the hypothesis of the present study can be stated as follows: Ingestion of LSD and dextroamphetamine is expected to cause an increased usage of novel figurative phrases by patients during psychoanalytic sessions.¹

Verbatim transcripts of three male psychoanalytic patients were used in the present study;

This research was aided in part by MH 13579, which supported the first author as a postdoctoral research fellow in the Interdisciplinary Research Training Program, Department of Psychiatry, Columbia University, and by MH 06470-19, which supported the second author as a psychology intern in the Department of Psychiatry, Columbia University.

The authors wish to express thanks to the William Alanson White Institute where the project was originally conducted.

Requests for reprints should be sent to Michael Natale, who is now at the Department of Psychology, University of Pennsylvania, 3813 Walnut Street, Philadelphia, Pennsylvania 19174.

¹ The interested reader is referred to Mechanek, Feldstein, Dahlberg, and Jaffe (1968) for a detailed description of the research project. What follows is a brief outline of the methodology.

the patients were white, college educated, and suffered from neurotic depression. All subjects provided informed consent to participate in the research project.

Each patient received each of the three drugs (LSD = 50–100 u; dextroamphetamine = 15 mg; placebo) seven times on a randomized schedule for a total of 21 experimental psychotherapy sessions over a period of 1½ years. Patients, therapist (CCD), and raters (MN and MK) were blind concerning the drug condition.

Each transcript was scored via Barlow, Kerlin, and Pollio's (1970) programmed scoring technique for the patients' use of figurative language by two raters. Fourteen categories of figurative language were scored, with each instance being classified as either "frozen" or novel. Frozen was defined as a figure of speech that is considered a part of common vocabulary. Novel was defined as a figure of speech that the rater judged to be unique in the context. The following dependent variables were obtained: total figurative phrases (FP), FP/1,000 words, novel figurative phrases (N), N/1,000, and N/FP%.

The two judges reached a high level of agreement on their scoring of FP (96%); an 88.9% agreement was obtained for classification of novel versus frozen. Only phrases that were mutually scored were used in the analysis. A statistical analysis was performed on each subject; LSD and dextroamphetamine sessions were individually compared to the placebo condition ($n = 7$ for each subject). Dunn's multiple comparison procedure was used to make the a priori comparison.

The results indicated that Patient 1 significantly increased ($p < .01$) his use of FP, FP/1,000 words, N, N/FP%, and N/1,000 words when under the influence of LSD. Patient 2 displayed LSD-induced increases in nonliteral language similar to the above-described findings for Patient 1, with the exception of FP and N/FP%

($p < .05$). In addition, FP/1,000 words was significantly increased ($p < .05$) for Patient 2 when under the influence of dextroamphetamine. Concerning Patient 3, the use of nonliteral language during psychoanalysis was not affected by either LSD or dextroamphetamine.

To sum up, the hypothesis that LSD increases the patient's use of novel figurative language in the course of psychoanalysis has been supported in two out of three case studies. However, it should be noted that the overall therapeutic value of psychotomimetics has not been demonstrated by this study. Although LSD does apparently increase the use of novel metaphors, and there is evidence that figurative language is associated with an important therapeutic process (insight), there is no evidence that drug-induced imagery will have the same beneficial effects as images that occur naturally.

References

- Abramson, H. A. (Ed.). *The use of LSD in psychotherapy and alcoholism*. New York: Bobbs-Merrill, 1967.
- Barlow, J., Kerlin, J. & Pollio, H. R. *Manual for the identification of figurative language* (Tech. Rep. 1). Knoxville: Metaphor Research Group, University of Tennessee, 1970.
- Billow, R. M. Metaphor: A review of the psychological literature. *Psychological Bulletin*, 1977, 84, 81–92.
- Fink, M. Brain function, verbal behavior, and psychotherapy. *Comprehensive Psychiatry*, 1974, 15, 257–266.
- Gottschalk, L. A., & Gleser, G. C. *Measurement of psychological states through content analysis*. Los Angeles: University of California Press, 1969.
- Mechanek, R., Feldstein, S., Dahlberg, C. C., & Jaffe, J. An experimental investigation of LSD as a psychotherapeutic adjunct. *Comprehensive Psychiatry*, 1968, 9, 490–498.

Received December 9, 1977 ■

Psychometric Correlates of the Mosher Forced Choice Guilt Inventory

Kevin E. O'Grady
University of Connecticut

Louis H. Janda
Old Dominion University

One-hundred one male and 135 female undergraduate psychology students were administered the Mosher Forced Choice Guilt Inventory (MFCGI), which consists of subscales measuring Sex Guilt, Hostility Guilt, and Morality-Conscience Guilt; the Repression-Sensitization Scale; the State-Trait Anxiety Inventory; the California F Scale; the Marlowe-Crowne Social Desirability Scale; and the Adult Nowicki-Strickland Locus of Control Scale. Inspection of correlations and principal component solutions for each sex indicated that guilt emerges as a construct separate from anxiety or authoritarianism. The applicability of the male form of the MFCGI for use with females is discussed.

The Mosher Forced Choice Guilt Inventory (MFCGI; Mosher, 1966) consists of 79 forced-choice items intended to measure three theoretically independent aspects of guilt: sex guilt (SG), hostility guilt (HG), and morality-conscience guilt (MCG). The purpose of the present research plan was twofold. First, an examination of the literature indicated the importance of assessing the relationship between the subscales of the MFCGI and several other personality inventories. Second, because the 1966 MFCGI, originally intended for use with males, has been used to measure guilt in females, an assessment of the factor structure in males and females of the 1966 MFCGI in relation to these other scales also seemed called for.

The MFCGI, together with the Repression-Sensitization (R-S) Scale; the California F Scale (F); the Adult Nowicki-Strickland Locus of Control (I-E) Scale; the State-Trait Anxiety Inventory (STAI), A-Trait form; and the Marlowe-Crowne Social Desirability Scale (M-C SDS) were completed by 236 undergraduate psychology students (101 males, 135 females). Separate principal components solutions for each sex were then obtained from the matrix of intercorrelations of the test scores. In each case three principal components with eigenvalues greater than 1.0 were retained and were rotated to

orthogonal simple structure using the normalized varimax method (see Table 1).

The results of the two principal components analyses were then compared using a technique developed by Schönemann and Carroll (1970). This procedure, a generalization of the orthogonal Procrustes problem, uses a least squares approach to fit one matrix to another. The overall degree of correspondence between the two matrices is provided by a measure of normalized symmetric error (ϵ). Fitting the male rotated component loadings to the female rotated component loadings indicated that the inventories were assessing the same underlying dispositions in the two groups ($\epsilon = .013$).

The first component, accounting for 31.2% of the total variance for the male sample, and 31.9% in the female sample, was labeled *admission of anxiety*. For males, the primary loadings ($\geq |\pm .5|$) consisted of the R-S (.892), F (.607), STAI (.852), and I-E (.674) scales. Primary loadings for females consisted of the R-S (.929), STAI (.892), I-E (.715), and M-C SDS ($-.548$) scales.

The second component, accounting for 26.9% of the total variance for the male sample and 25.7% in the female sample, was almost entirely marked by the three guilt subscales of the MFCGI and clearly represents a general guilt factor. Male loadings were .704 for SG, .850 for HG, and .846 for MCG. Female loadings were .790 for HG, and .857 for MCG, and .425 for SG.

The third component, accounting for 13.5% of the total variance for the male sample and 15.0% in the female sample, seemed to be largely determined by authoritarianism in both

Thanks are due to Eleanor Wertheim, Charlene Alling, Dave Henry, Suzi Zlotlow, and Don Mosher for their comments on an earlier draft of this article.

Requests for reprints and for an extended report of this study should be sent to Kevin E. O'Grady, Department of Psychology, University of Connecticut, Storrs, Connecticut 06268.

Table 1

Means, Standard Deviations, and Intercorrelations of the Eight Scales for the Male and Female Samples

Variable	1	2	3	4	5	6	7	8
1. R-S		.39	.76	.46	-.35	-.10	.09	.00
2. F	.24		.37	.32	.07	.21	.16	-.09
3. STAI	.82	.11		.39	-.24	-.09	.20	.07
4. I-E	.56	.27	.51		-.15	.01	.03	.04
5. M-C SDS	-.41	.06	-.35	-.17		.27	.23	.09
6. SG	-.06	.39	-.01	.04	.25		.61	.38
7. MCG	.06	.16	.06	-.01	.31	.51		.52
8. HG	.11	-.09	.13	.03	.16	.19	.45	
Males								
M	39.60	11.04	34.69	9.38	15.39	10.53	11.86	17.54
SD	17.80	4.91	8.47	3.87	5.54	7.08	4.58	5.86
Females								
M	43.76	10.82	37.50	10.20	15.93	12.81	13.44	19.52
SD	17.57	5.21	8.24	4.95	5.49	6.61	3.94	4.52

Note. R-S = Repression-Sensitization Scale; F = Authoritarianism Scale; STAI = State-Trait Anxiety Inventory; A-Trait Form; I-E = Internal-External Locus of Control Scale; M-C SDS = Marlowe-Crowne Social Desirability Scale; SG = Sex Guilt subscale, MCG = Morality-Conscience Guilt subscale, HG = Hostility Guilt subscale of the Mosher Forced Choice Guilt Inventory. The correlation for males are above the diagonal, and the correlations for females are below it. For males, $r(99) \geq .19$, $p < .05$; $r(99) \geq .25$, $p < .01$. For females, $r(133) \geq .17$, $p < .05$; $r(133) \geq .23$, $p < .01$.

groups. Male loadings were composed of the F (.644) and M-C SDS (.707) scales, along with a smaller loading for SG (.453). Female loadings primarily involved the F (.857) and SG (.717) scales.

Examination of the separate factor solutions, and the residual matrix (the matrix of differences between the matrix of best fit and the initial target matrix) provided through the Procrustes procedure, revealed that the lack of correspondence in the solutions was mainly due to the fact that the SG subscale loaded primarily on the second factor for males, with a secondary loading on the third factor, whereas this pattern was reversed for the female subjects.

These results generally support the notion that guilt is a construct identifiably different from anxiety or authoritarianism. Besides a nominal amount of shared variance with SD for SG ($r = .25$ for males, $r = .27$ for females) and MCG ($r = .23$ for males, $r = .31$ for females, all $ps < .01$), only three other significant relationships emerged that involve the subscales of the MFCGI: SG with F for males ($r = .21$; $p < .05$) and females ($r = .39$, $p < .01$), and MCG with STAI for males ($r = .20$, $p < .05$). In contrast, the three subscales of the MFCGI share a moderate amount of variance among themselves (ranging from 4% to 38% across the two groups). Together with the results of the multi-trait-multimethod analysis of Mosher (1966),

which indicated that the SG, HG, and MCG subscales of the MFCGI measure discriminately different aspects of guilt, this study, which indicates that guilt is a discriminately different construct from anxiety or authoritarianism, offers strong support for the construct validity of the MFCGI.

Conclusions regarding the use of the male form of the MFCGI with females remain tenuous. Although the intercorrelations of the three subscales of the MFCGI exhibited the same overall pattern in males and females, the female correlations were somewhat lower than the corresponding male correlations. Further, the difference in factor structure for SG in the male and female factor solutions suggests that the SG subscale may be measuring different dispositions in the two groups. A comparison of the item factor structure of the MFCGI between males and females would help to clarify this issue.

References

- Mosher, D. L. The development and multitrait-multimethod matrix analysis of three measures of three aspects of guilt. *Journal of Consulting Psychology*, 1966, 30, 25-29.
- Schönemann, P., & Carroll, R. Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, 1970, 35, 245-255.

Received December 15, 1977 ■

Clinical Application of a Broad-Spectrum Behavioral Approach to Chronic Smokers

Harry A. Lando and John A. McCullough
Iowa State University

The present investigation assessed the clinical applicability of a broad-spectrum behavioral treatment (aversion, contractual management, booster sessions, group contact and support) previously reported by Lando. The feasibility of reorienting non-abstinent subjects to maintained reduction was also assessed. Twelve of 16 subjects who completed treatment were abstinent at a 6-month follow-up (75%), and 2 others were smoking substantially less than at baseline. Results suggest that the previous findings are replicable and that this program can be applied on a self-supporting clinical basis. Additional work is needed to validate the maintained reduction procedure.

Lando (1977) has called for the application of promising laboratory approaches to smoking cessation on a clinical basis. In that article he described a broad-spectrum treatment consisting of aversion, contractual management, booster sessions, faded group contact, and support, which led to a 76% abstinence rate at a 6-month follow-up. The major purpose of the present study was to assess the applicability of this treatment in a cost-effective clinical program. The present study was also intended to provide a replication of the Lando (1977) findings, as well as to allow a preliminary assessment of an additional treatment component in which nonabstinent smokers are reoriented toward maintained reduction.

Treatment was conducted by the second author, who is an advanced psychology graduate student at Iowa State University. He followed the treatment procedures for aversion and maintenance outlined in a treatment manual published by Lando (1976). Participants were required to deposit \$20, of which \$10 was refundable. The remaining \$10 was sufficient to cover all costs of the program.

Subjects included 17 smokers. (M age was 35.7 years, M years as a smoker was 17.3, and M baseline smoking was 28.7 cigarettes per day.)

This investigation was supported by Grant 1 R01 CA22102-01, awarded by the National Cancer Institute, Department of Health, Education, and Welfare.

Requests for reprints should be sent to Harry A. Lando, Department of Psychology, Old Botany Hall, Iowa State University, Ames, Iowa 50011.

Procedures for record keeping were as described in previous studies.

Aversion included six treatment sessions over a 1-week period. In the current study all 17 subjects were seen in a single group. Sessions averaged approximately 45 minutes, including a prescribed 25-minute period of continuous smoking. Subjects were also instructed to smoke as much as possible on their own between sessions, with a daily minimum of twice their usual consumption.

Subjects were instructed to abstain following aversion, and they attended seven maintenance sessions at the termination of this phase of the program. Formal support was faded over time, with the initial session occurring within 48 hours and the second session taking place 5 days later. Three additional maintenance sessions were then held at weekly intervals, followed by two final sessions that took place 2 weeks apart. These sessions consisted primarily of group discussion and the signing of contracts. Subjects contracted to forfeit money for every cigarette smoked, to reward themselves for increasingly longer periods of abstinence, and to undergo 48 hours of booster aversion (rapid smoking) in the event of relapse. (A more detailed description of treatment is available elsewhere; Lando, 1976, 1977.)

In a departure from previous work, subjects who failed to remain abstinent following two booster treatments were reoriented toward maintained reduction. There appear to be a certain number of smokers for whom abstinence is virtually impossible. In the current study maintained reduction was explored as a possible alternative for these smokers. The same essential

contractual techniques that had been directed toward abstinence were adapted to a program of maintained reduction. Nonabstinent subjects contracted to limit their smoking to a set daily allotment (between 10 and 20 cigarettes). Subjects were maintained at this level rather than gradually reduced to minimize the likelihood that each remaining cigarette would become more reinforcing. Efforts were also made to minimize the reinforcing value of smoking by prescribing a stimulus control procedure in which smoking was confined to specified locations and was divorced from other rewarding activities. Subjects now forfeited money for every cigarette smoked *over* the daily allotment and rewarded themselves over increasingly longer intervals for adhering to targeted consumption.

Sixteen subjects completed treatment, with 1 other dropping out at the termination of aversion. Of the subjects who completed the program, 100% remained abstinent at Week 1. Two subjects relapsed during the 1st month following aversion, and 2 others reported relapse in the 4th month. This yielded a 2-month abstinence rate of 88% (14 of 16 subjects), and 4- and 6-month abstinence rates of 75% (12 of 16 subjects). If the person who dropped out of the program is counted as a treatment failure, the 6-month abstinence rate is 71%.

The two subjects who relapsed during the first month failed to respond to booster treatment and were placed on maintained reduction (12 subjects maintained abstinence for the entire 6 months and did not undergo booster aversion). The 2 subjects who relapsed during the 4th month had reverted to their original levels of smoking by the 6-month follow-up. (No pressure was placed on subjects to participate in booster sessions or maintained reduction following the 2-month maintenance program.) Both subjects who were assigned to maintained reduction contracted to limit their smoking to 20 cigarettes a day. (Both had previously been two-pack-a-day smokers.) One of these subjects averaged 13 cigarettes per day at the 4-month follow-up and 17 at the 6-month follow-up (more than 5 months after the maintained reduction program had been initiated). The other subject averaged 21 and 25 cigarettes per day at the 4- and 6-month follow-ups, respectively. Unfortunately, the appropriate contingencies were not fully implemented, with contractual responsibility left to subjects and apparently only loosely adhered to.

Overall, the present findings are extremely encouraging, primarily because they suggest that the successful program reported by Lando (1977) is both replicable and generalizable to a clinical context. Treatment was offered on a service basis using a programmed treatment manual. This was not a controlled laboratory study, nor was it intended to be. Rather, it was an exploratory assessment of the applicability of a cost-effective laboratory-based program as a self-supporting clinical treatment. By this criterion, the program was clearly successful: The results are comparable to those obtained by Lando (1977) and are among the most encouraging findings reported in the literature to date.

Caution is still indicated in interpreting these findings, however. Both the current study and that of Lando (1977) were based on relatively small numbers of subjects. Both studies were conducted in the same community, with a similar population of smokers. There is considerable need for replication of this program by other investigators in different settings. The fact that all subjects were seen in a single treatment group also limits the generalizability of the current findings.

Considerable group cohesiveness was again evident in the present program and may be due at least in part to the specific treatment techniques used. Subjects were repeatedly informed of the importance of their own efforts in aiding themselves and in supporting each other. Participation in the program itself represented a significant commitment both in undergoing the aversion and maintenance sessions and in remitting the \$10 fee. The commitment to the program (and to each other) seemed to be reinforced by the extremely encouraging initial outcome in which 100% of the participants remained abstinent throughout Week 1.

Another purpose of the present study was to assess the effectiveness of a maintained reduction program for nonabstinent subjects. Although such a program appears very promising, particularly in providing nonabstinent subjects with an alternative, it cannot be meaningfully evaluated on the basis of the current evidence. Only two subjects underwent maintained reduction, and although the program appeared to be at least moderately successful in both cases, the contractual and stimulus control procedures were not fully implemented. The possibility that these subjects might eventually resume their former levels of smoking must also be considered. In future work maintained reduction programs should be more carefully implemented, and care

Catalog of Selected Documents in Psychology,
1976, 6, 113 (Ms. No. 1371).

Lando, H. Successful treatment of smokers with a broad-spectrum behavioral approach. *Journal of Consulting and Clinical Psychology*, 1977, 45, 361-366.

References

Lando, H. A manual for a broad-spectrum behavioral approach to cigarette smoking. *ISAS*

Received December 16, 1977 ■

[illegible]

Relationships Between Client Self-Perceptions of Self-Consciousness Levels and Therapist Awareness of These Perceptions

Robert G. Turner and Mae Keyson
Pepperdine University

The present study investigated the congruity between clients' self-perceptions related to self-consciousness and their therapists' awareness of these self-perceptions. Neurotic ($n = 47$) and psychotic ($n = 51$) clients completed the Self-Consciousness Scale (SCS) under standard self-descriptive instructions. Therapists completed the SCS as they thought their client would respond. In both samples of clients, therapist's reports were significantly correlated with client self-reports on subscales of the SCS indicating self-reflective habits (Private Self-Consciousness) and discomfort in the presence of others (Social Anxiety) but not on a measure of awareness of oneself as a social object (Public Self-Consciousness).

The Self-Consciousness Scale (SCS; Fenigstein, Scheier, & Buss, 1975) is a measure of individual differences in habitual self-focused attention. Factor analysis of the items in the scale revealed two components of self-consciousness. The factor labeled *Private Self-Consciousness* concerns habitual attendance to one's thoughts, motives, and feelings and thus measures self-reflective or introspective tendencies. High compared to low private self-consciousness subjects generate self-reports that are both more detailed (Turner, in press) and more predictively valid (Turner, 1978). The second factor, Public Self-Consciousness, was defined by a general awareness of the self as a social object and reflects a concern for social appearance. A third factor, Social Anxiety, also emerged from the analyses of the scale and reflects a measure of discomfort in the presence of others. Social anxiety presumably occurs as a possible reaction to public self-consciousness. (See Turner, Scheier, Carver, & Ickes, 1978, for a review of research involving these subscales.)

Fenigstein et al. (1975) argued for the importance of individual differences in client self-consciousness levels in determining both the choice and goals of therapy. The goals of therapy might be directed toward the modification of a component of self-consciousness, for ex-

ample, decreased social anxiety or increased insight and self-reflection.

A component of many therapeutic approaches consists of the therapist understanding the phenomenological world of the client. Since self-consciousness would seem to have broad clinical and behavioral implications (cf. Turner et al., 1978), it would be useful to know the extent to which therapists are aware of clients' self-perceptions as related to self-consciousness levels. For example, if the goal of a therapeutic program is to increase client insight, are therapists aware of the extent to which their clients consider themselves to be self-reflective or insightful? Therefore, the purpose of the present investigation was to determine the congruity between clients' self-perceptions related to self-consciousness and their therapist's awareness of these self-perceptions.

The sample of clients consisted of 47 outpatients of a private psychiatric center who were clinically diagnosed as psychoneurotic disorder and 51 resident patients of a private, non-profit psychiatric hospital who were clinically diagnosed as psychotic disorder (schizophrenic reaction). Each client was paired with the therapist directly in charge of administering the therapeutic program of the client. Part of the theoretical orientation of each therapist emphasized understanding the client's phenomenological world. There were four therapists directing programs for the neurotic sample (each having 8-17 clients) and seven directing programs for the psychotic sample (each having 3-13 clients).

Each client completed the SCS under standard self-descriptive instructions. Independently, ther-

The assistance of Sandra Burnett is gratefully acknowledged.

Requests for reprints should be sent to Robert G. Turner, Social Science Division, Pepperdine University, Malibu, California 90265.

Table 1
Correlations Between Self-Reports and Therapists' Reports of Clients' Scores on the Self-Consciousness Subscales

Therapists' reports	Self-reports					
	Neurotic sample (<i>n</i> = 47)			Psychotic sample (<i>n</i> = 51)		
	1	2	3	1	2	3
1. Private Self-Consciousness	.34*	.28	-.01	.46***	.38**	.10
2. Public Self-Consciousness	.05	.27	.40**	.12	.25	.08
3. Social Anxiety	.06	.24	.54***	.24	.33*	.43**

* $p < .05$.

** $p < .01$.

*** $p < .001$.

apists were instructed to "answer the items as you think the client would respond."

Table 1 presents correlations between self-reports and therapists reports of the scores of clients on the SCS subscales. In both samples the correlations of primary interest lie along the major diagonal. Completing the SCS as they thought their clients would, therapists generated subscale scores significantly correlated with the self-reports of their clients for private self-consciousness and social anxiety in both samples of clients. Thus therapists were significantly aware of the extent to which clients would report themselves to be self-reflective and uncomfortable as the object of public attention.

In both samples of clients, therapists' reports and self-reports on the Public Self-Consciousness subscale were not significantly correlated. To the therapists of the neurotic sample, public self-consciousness was related to client social anxiety levels. In the psychotic sample, client perceptions tended to be related to therapists' perceptions of the clients as self-reflective or socially anxious. Thus the therapists' reports of clients' public self-consciousness were neither significantly nor distinctively related to clients' self-reports of public self-consciousness.

The level of the off-diagonal correlations of public self-consciousness with private self-consciousness and social anxiety are consistent with intercorrelations of these subscales in single-sample data (Turner et al., 1978). It would appear that public self-consciousness is neither as salient nor as relevant to insight-based therapeutic processes as are private self-consciousness and social anxiety. Self-reflection is a primary goal of such approaches, and levels of social anxiety are immediately salient in group therapy

procedures that often are a part of insight therapeutic programs. Concern for how one appears to others (public self-consciousness) would not be relevant unless it resulted in discomfort in the presence of others or social anxiety. Thus normal levels of public self-consciousness would not appear to be particularly relevant in many insight therapeutic programs.

If the therapist is to facilitate the growth of self-insight in the client, awareness of self-perceptions that are relevant to the direction of therapy is imperative. A possibly fruitful area for future research would be to investigate the relationships between client self-perceptions and therapists' awareness of these perceptions in other important areas of client self-concept. Indications of common accurate and inaccurate perceptions might assist therapists in identifying areas requiring diligent attention if accurate client self-perceptions are to be achieved.

References

- Fenigstein, A., Scheier, M., & Buss, A. H. Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology*, 1975, 43, 522-527.
- Turner, R. G. Consistency, self-consciousness, and the predictive validity of typical and maximal personality measures. *Journal of Research in Personality*, 1978, 12, 117-132.
- Turner, R. G. Effects of differential request procedures and self-consciousness on trait attributions. *Journal of Research in Personality*, in press.
- Turner, R. G., Scheier, M. F., Carver, C. S., & Ickes, W. Correlates of self-consciousness. *Journal of Personality Assessment*, 1978, 42, 285-289.

Received December 16, 1977 ■

Situational Management, Standard Setting, and Self-Reward in a Behavior Modification Weight Loss Program

Stanley L. Chapman

Department of Rehabilitation Medicine
Emory University

D. Balfour Jeffrey

University of Montana

In a comprehensive weight loss program, overweight women exposed to instruction in self-standard setting as well as to situational management techniques lost more weight than those instructed only in situational management techniques. Addition of instruction in self-reinforcement to standard setting and situational management failed to produce additional weight loss. Findings illustrate the facilitative effect of teaching individuals to set specific, objective, and realistic goals for eating behaviors and weight and the difficulties of incorporating self-reward procedures in a comprehensive program.

Short-term behavioral programs have yielded promising results, but there has been insufficient research on the variables involved (Abramson, 1977). Although several investigators have found greater weight losses in individuals taught to reward themselves for changing eating habits or weight than in controls instructed to self-monitor eating and weight (e.g., Bellack, 1976; Mahoney, 1974), the role of standard setting as a component of self-reinforcement has not been adequately studied.

In the present study, 57 women, who ranged from 17.0% to 84.8% above ideal body weight ($M = 45.4$) on insurance tables were recruited from the community through newspaper ads. They ranged in age from 17 to 65 years ($M = 37.8$), and most came from a middle or upper-middle socioeconomic level. They deposited \$40, which was returned contingent on attendance at group meetings, and were divided into triplets of approximately equal body weights and percentages of overweight. Members of each triplet were assigned randomly to the three treatment groups, each of which met for 90 minutes for 8 consecutive weeks and for a follow-up meeting 8 weeks after treatment. Group meetings consisted of a private weigh-in and presentation and discussion of specific behavioral techniques to be used during the following week. Three therapists, all of whom were familiar with behavior modification but who were in-

experienced in behavioral treatment for weight, rotated as leaders for the three groups. They were unaware of the hypotheses of the study and did not set standards or reward group members verbally for any eating behaviors or weight losses.

Treatment groups included a situational management group (SM), which focused mainly on the control of environmental stimuli that lead to overeating, as outlined by Ferguson (1975). Instructions in self-monitoring, stimulus control techniques, chaining, exercise, and diet were included. A second group, the self-standard setting group (SS), received the same instructions as Group SM, but they also were taught to write down standards each week for changes in weight and eating behaviors. SS participants were encouraged to set goals that they had an excellent chance to achieve and to make them specific, such as "I will eat a moderate breakfast (below 500 calories) each morning," rather than "I will never eat candy again." (See Jeffrey & Katz, 1977, for more information on realistic standard setting and other behavioral techniques.) Therapists stressed the importance of standards and examined each participant's standards weekly.

The self-reward (SR) group received the same training as Groups SM and SS, but members also were instructed weekly to make contracts with themselves and others to receive verbal and material rewards contingent on successful weight losses and small daily achievements in changing eating behaviors. Suggested examples included telling oneself one performed well, arranging small treats such as a leisurely bath or a trip to the beauty parlor, and arranging a special outing with one's spouse.

In addition, members of this group deposited an extra \$17.50, which was used as monetary

Thanks are extended to Patricia Chapman, Alan Torppa, and Patricia Weis, who served as therapists for this study.

Requests for reprints and for an extended report of this study should be sent to D. Balfour Jeffrey, Department of Psychology, University of Montana, Missoula, Montana 59812.

rewards for improvement in the form of a check for \$1.25 for eating habit changes and a check for \$1.25 for weight loss each week for 7 weeks of treatment. Participants were instructed to reward themselves weekly by taking their respective checks only if they reached their respective weekly goals for eating and weight loss. If they did not reach their goals, they were instructed *not* to reward themselves, but rather to deposit the checks through a slot in a locked box outside the meeting room.

One member of each group dropped out of treatment, and one SR group member was excluded from follow-up analyses after developing thyroid difficulties. Mean weight changes were compiled for each group for a 7-week baseline period, for 7 weeks of treatment, and for 1 follow-up 8 weeks after treatment. These results were as follows: Group SM ($n=19$) gained 1.84 pounds (.83 kg) during baseline, lost 4.03 pounds (1.82 kg) during treatment, and lost an additional 1.13 pounds (.51 kg) after treatment; Group SS ($n=18$) gained .16 pounds (.07 kg) during baseline, lost 7.47 pounds (3.39 kg) during treatment, and lost 1.97 pounds (.89 kg) after treatment; Group SR ($n=17$) gained 1.57 pounds (.71 kg) during baseline, lost 4.71 pounds (2.14 kg) during treatment, and lost 1.46 pounds (.66 kg) after treatment. One-way analyses of variance with repeated measures indicated that all three groups lost significant amounts of weight during the treatment and from pretreatment to follow-up (all p 's $< .01$). Welch's t test revealed that Group SS lost significantly more weight than Group SM during treatment and from the beginning of treatment to the end of follow-up ($p < .05$), but all other group comparisons were nonsignificant.

At posttreatment and at follow-up, members of each group rated the helpfulness of 19 behavior change techniques. "Setting standards for behavior" received the highest mean rating in the SS group and also was rated highly in Group SR, whereas Group SR rated "rewarding self" and "getting others to reward me" as the two least helpful techniques that they learned. All correlations calculated between weight change during and after treatment and self-reports of standard setting for eating habits reached the .05 level of significance for the SS and SR groups combined. These coefficients, which ranged from $-.343$ to $-.609$, were calculated from self-reports on the closeness of attention paid to eating standards, the number of behaviors for which standards were set, the frequency of setting standards for eating, and the success in reaching them. Closeness of attention paid to weight standards correlated significantly

with weight loss after treatment but not during treatment. Questionnaire data also showed that SR participants gave themselves a mean of only two verbal rewards per week at home and one material reward every 3-4 weeks.

The significant weight loss of the SM group adds further evidence that weight loss can be produced without programming external or self-reinforcement. The data also support the superiority of teaching individuals to set standards for eating and for losing weight in addition to teaching situational management over instruction in situational management alone. On the other hand, comprehensive instruction in self-reward in addition to standard setting and situational management did not facilitate further weight loss as predicted. This instruction may have been incompatible with instruction in the standard setting or it may have caused participants to be overloaded with information. However, these possibilities were not consistent with questionnaire and anecdotal data, which indicated that the self-reward techniques were seen as being of little benefit, were used seldom, and made participants feel anxious and overly pressured to achieve standards. Many of them reported that losing weight was so important to them anyway that instruction in self-reward was unnecessary.

It is possible that implicit training in setting standards rather than training in self-reinforcement was the more important factor in previous studies in which members of self-reward groups have shown greater weight losses than controls. External or self-standard setting was embedded in the self-reward condition but not in the control conditions of these studies. The findings of this study do not imply that self-reinforcement is unimportant, but they illustrate the difficulties of incorporating self-reward procedures in a comprehensive clinical program.

References

- Abramson, E. E. Behavioral approaches to weight control: An updated review. *Behaviour Research and Therapy*, 1977, 15, 355-363.
- Bellack, A. S. A comparison of self-reinforcement and self-monitoring in a weight reduction program. *Behavior Therapy*, 1976, 7, 68-75.
- Ferguson, J. M. *Learning to eat: Behavior modification for weight control*. Palo Alto, Calif.: Bull, 1975.
- Jeffrey, D. B., & Katz, R. C. *Take it off and keep it off: A behavioral program for weight loss and healthy living*. Englewood Cliffs, N.J.: Prentice-Hall, 1977.
- Mahoney, M. J. Self-reward and self-monitoring techniques for weight control. *Behavior Therapy*, 1974, 5, 48-57.

Received December 19, 1977 ■

Psychotherapeutic Styles: Self-Perceptions of Therapists Varying on A-B Type and Experience

William B. Goodwin
Yale University

The possibility that psychotherapeutic styles might differ as a function of A-B type or experience was investigated by interviewing 40 A and B experienced and inexperienced psychotherapists on a wide range of issues relating to therapeutic approach and style. The importance of the experience level of the therapist was confirmed by therapists' self-descriptions: Experienced therapists seemed to conform less to conventional stereotypes of the traditional psychotherapist than did inexperienced therapists, and experienced therapists appeared to emphasize immediacy and encounter while deemphasizing detachment, diagnoses, and psychodynamic interpretations. At the same time, self-perceived differences between A and B therapists were not apparent.

One approach to the study of psychotherapy that has seldom been tried is to interview the therapists themselves. Schiffman, Carson, and Falkenberg (Note 1) A-B scale scores were obtained by mail from a large sample of trainee therapists at the Yale University Department of Psychiatry. A-B scores were also available for a number of experienced therapists who had participated in a previous study (Goodwin, 1975). Only male clinicians with at least 1 year of therapeutic experience and who were clearly Type A or Type B as defined by the approximate outer quartiles of the distribution (14 or above = A; 8 or below = B) were invited to participate. Forty of the 41 therapists who were contacted consented to be interviewed about their own approach to psychotherapy, in exchange for a précis of the results upon completion of the study. Each clinician was then scheduled for a 1-hour standardized tape-recorded interview in his own office, followed by a brief questionnaire to elicit extent of use of particular techniques.

All scales and coding systems for interview responses were constructed after preliminary inspection of the range of responses. The interviewer (myself) and corater blindly rated the

material. I was blind to therapists' A-B type until the coding was completed.

Results

Interscorer reliabilities for all response scales and category systems were quite high, with many in the .90s. Mean A-B scores were 16.75 for As ($n = 20$, $SD = 1.92$) and 6.05 for Bs ($n = 20$, $SD = 1.40$). Therapists still in training averaged 2.10 years of experience; experienced therapists averaged 9.95 years of experience.

Outcome Index

For responses to four questions regarding retrospective assessment of psychotherapy outcome, an outcome index was devised based on a simplified version of the A-B interaction hypothesis: Positive outcome with schizoids or negative outcome with neurotics was scored 3; positive outcome with neurotics or negative outcome with schizoids was assigned 1; and other responses were assigned 2. The questions addressed the individual patient with whom the therapist had enjoyed the greatest success, the one with whom he had encountered the least success, and the type of patient with whom he had experienced the greatest and least success. Scores for these questions were summed, and the resulting linear combination was subjected to a 2×2 analysis of variance. No main effect for A-B type was found nor were other effects significant.

Factor Analysis

The bulk of the responses were subjected to a principal components factor analysis; varimax rotations were then computed for a four-factor

This article is based on a doctoral dissertation submitted to the Department of Psychology at Yale University under the supervision of Donald M. Quinlan and Jesse D. Geller.

Requests for reprints, an extended report including the factor loadings matrix, and copies of the interview schedule should be sent to William B. Goodwin, 599 Whitney Avenue, Apartment 1A, New Haven, Connecticut 06511.

solution. Factor labels, and proportions of variance accounted for, were 1: Human Potential Movement (18.4%); 2: Rogerian Style (14.1%); 3: Modified Psychoanalytic Style (13.1%); and 4: Agitated Involvement Versus Calm Neutrality (11.4%). A 2×2 multivariate analysis of variance was computed on the four factor scores for each therapist, treating A-B type and experience level as the predictors. Neither the overall A-B Type \times Experience Level interaction nor the A-B main effect were significant. Thus, a pervasive difference in style between A and B psychotherapists was not evident.

Experience Level

The multivariate main effect for experience level was significant ($p < .017$), however, and chiefly reflected experienced therapists' low scores on the Modified Psychoanalytic Style factor (weighted $-.905$ in the discriminant function) and high scores on the Human Potential Movement factor (weighted $.514$ in the discriminant function). Based on the most salient components of these two factors—and noting univariate F s where significant—inexperienced therapists seemed to perceive themselves as rarely using an advice-giving educational approach, $F(1, 16) = 4.73$, $p < .036$; as not personally involved with, $F(1, 16) = 6.15$, $p < .018$, or self-disclosing toward their patients; as making extensive use of psychodynamic interpretations, $F(1, 16) = 6.31$, $p < .017$; as seeing diagnoses as very useful, $F(1, 16) = 7.03$, $p < .012$; as not expressing negative reactions readily to patients, $F(1, 16) = 4.82$, $p < .035$; and as being inclined to reject the "sociopathology of mental illness" viewpoint of Thomas Szasz.

Experienced therapists perceived themselves as not using depth interpretations frequently; as tending to use an advice-giving educational approach; as being more personally involved with and self-disclosing toward their patients; as expressing negative reactions readily; as using bantering to a moderate extent; as finding diagnoses comparatively useless; and as tending to agree with the viewpoint of Thomas Szasz.

Possible Covariates

No significant differences between As and Bs were found with respect to type of practice, place of training, avowed theoretical orientation, or profession, but inexperienced therapists more frequently had institutional practices (100% vs. 60%) and were more frequently trained at Yale University than their experienced colleagues (100% vs. 30%). In addition, a higher

proportion of the inexperienced (85%) than of the experienced therapists (40%) were psychiatrists. The possibility that experience-level differences were actually mediated by profession differences was indirectly evaluated by means of a profession by experience level 2×2 multivariate analysis of variance of the factor scores: No main or interaction effects for profession were found, however.

Discussion

Perhaps broad differences in clinical style between A and B therapists do not exist, although it remains possible that such differences may not be reflected in therapist self-description or may not be aligned along dimensions studied here.

Perhaps the most striking thing about the self-descriptive portraits for the two experience levels is that popular stereotypes coincide more closely with the self-characterizations of the inexperienced therapists. It could be that one of the major effects of accruing therapeutic experience is a moving away from traditional role expectations toward higher degrees of personal involvement and encounter with patients. Longitudinal studies of professional psychotherapists are obviously needed to resolve this question.

Finally, if such differences between experienced and trainee therapists can be replicated in other populations using other methods, what is it about therapist experience level that creates these differences? Maturation, aging, increased life experiences, change in socioeconomic status, and possible additional training are a few alternatives to the obvious increase in psychotherapy experience. By partitioning experience level into more subgroupings, future research might also attempt to ascertain whether any such differences emerge gradually and continuously or abruptly and discontinuously.

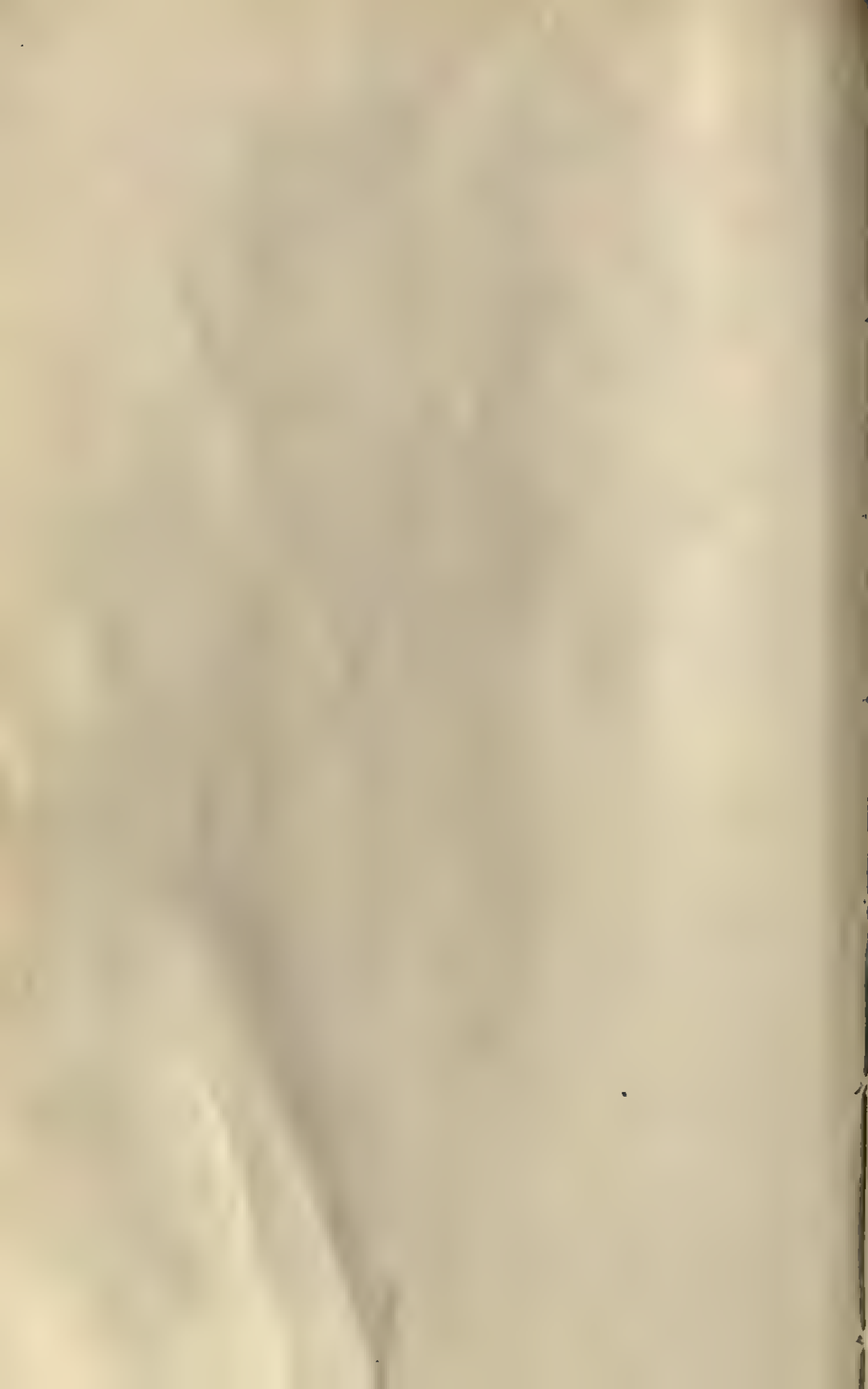
Reference Note

1. Schiffman, H., Carson, R. C., & Falkenberg, P. *A psychometric analysis of the Whitehorn-Bels A-B scale*. Unpublished manuscript, Duke University, 1967.

Reference

- Goodwin, William B. The A-B variable and stylistic differences in psychotherapy. (Doctoral dissertation, Yale University, 1975). *Dissertation Abstracts International*, 1975, 36, 2470B-2471B. (University Microfilms No. 75-24,537).

Received August 1, 1977 ■



Journal of Consulting and Clinical Psychology

Volume 46, 1978

Brendan A. Maher, Editor
Harvard University

Judith P. Worell, Associate Editor
University of Kentucky

Published bimonthly by the
American Psychological Association, Inc.
1200 Seventeenth Street, N.W., Washington, D.C. 20036
Copyright © 1978 by the American Psychological Association, Inc.

Consulting Editors

- Thomas M. Achenbach, *National Institute of Mental Health, Bethesda, Maryland*
 Kent G. Bailey, *Virginia Commonwealth University*
 Barbara B. Brown, *Beverly Hills, California*
 James N. Butcher, *University of Minnesota*
 Jean Chapman, *University of Wisconsin—Madison*
 George M. Chartier, *Arizona State University*
 Andrew L. Comrey, *University of California, Los Angeles*
 W. Grant Dahlstrom, *University of North Carolina, Chapel Hill*
 Anthony Davids, *Brown University*
 Gerald C. Davison, *State University of New York at Stony Brook*
 Alan S. DeWolfe, *Loyola University of Chicago*
 Barbara Dohrenwend, *Graduate School and University Center of the City University of New York*
 Richard M. Eisler, *Virginia Polytechnic Institute and State University, Blacksburg, Virginia*
 Norman S. Endler, *York University, Downsview, Canada*
 Sol Garfield, *Washington University in St. Louis*
 Donna M. Gelfand, *University of Utah*
 Goldine C. Gleser, *University of Cincinnati Medical Center*
 Harrison G. Gough, *University of California, Berkeley*
 Jesse G. Harris, Jr., *University of Kentucky*
 Mary B. Harris, *University of New Mexico*
 Kenneth Heller, *Indiana University*
 Frances A. Hill, *University of Montana*
 Alexandra G. Kaplan, *University of Massachusetts—Amherst*
 Alan S. Kaufman, *University of Georgia*
 Alan E. Kazdin, *Pennsylvania State University*
 Jeffrey A. Kelly, *University of Mississippi Medical Center*
 Barbara K. Keogh, *University of California, Los Angeles*
 James C. Kincannon, *University of Minnesota*
 Richard I. Lanyon, *Arizona State University*
 Herbert M. Lefcourt, *University of Waterloo, Waterloo, Canada*
 Peter M. Lewinsohn, *University of Oregon*
 Michael J. Mahoney, *Pennsylvania State University*
 G. Alan Marlatt, *University of Washington*
 Albert Marston, *University of Southern California*
 Eric J. Mash, *University of Calgary, Calgary, Canada*
 Richard McFall, *University of Wisconsin—Madison*
 Martha T. S. Mednick, *University of Connecticut*
 Barbara G. Melamed, *University of Florida*
 David I. Mostofsky, *Boston University*
 Benjamin J. Murawski, *Peter Bent Brigham Hospital, Boston, Massachusetts*
 Peter E. Nathan, *Rutgers—The State University*
 John M. Neale, *State University of New York at Stony Brook*
 K. Daniel O'Leary, *State University of New York at Stony Brook*
 Oscar A. Parsons, *University of Oklahoma*
 Nolan E. Penn, *University of California, San Diego*
 N. Dickon Reppucci, *University of Virginia*
 Jerome Resnick, *Temple University*
 Leonard G. Rorer, *Miami University*
 Gerald M. Rosen, *Providence Family Medical Center, Seattle, Washington*
 Robert Rosenthal, *Harvard University*
 William Schofield, *University Hospitals, Minneapolis, Minnesota*
 Carolyn Simmons, *University of Colorado at Denver*
 Bonnie Strickland, *University of Massachusetts—Amherst*
 Hans H. Strupp, *Vanderbilt University*
 Richard M. Suinn, *Colorado State University*
 Richard S. Surwit, *Duke University Medical Center*
 Patricia B. Sutker, *Medical University of South Carolina*
 Norman F. Watt, *University of Massachusetts—Amherst*
 Robert L. Weiss, *University of Oregon*
 Joan Welkowitz, *New York University*
 G. Terence Wilson, *Rutgers—The State University*
 Patricia Wisocki, *University of Massachusetts—Amherst*
 Melvin Zax, *University of Rochester*

Acknowledgment

In addition to the regular panel of Consulting Editors, the following people reviewed and evaluated manuscripts for the *Journal of Consulting and Clinical Psychology* during the preparation of the 1978 volume.

Allen Adinolfi
James Alexander
Norman H. Anderson
Stephen Auerbach

Bruce L. Baker
Robert F. Bales
Joseph Becker
Irving Beiman
Alan S. Bellack
Herbert Benson
Peter M. Bentler
Allen Bergin
Douglas A. Bernstein
Ellen Berscheid
J. Allen Best
Anthony Biglan
Henry Biller
Robert C. Birney
George E. Bigelow
Gary R. Birchler
Edward B. Blanchard
Paul H. Blaney
Sidney J. Blatt
Jeanne H. Block
Marcia M. Bok
Philip Bornstein
Stanley Brodsky
Denis B. Bromley
Inge K. Broverman
Fred G. Brown
Brenda Bryant
Elaine P. Burgess
Thomas G. Burish
Arnold H. Buss
Nelson Butters
Billy Barios

Dale Callner
Bonnie W. Camp
Mary M. Campbell
Albert S. Carlin
Carol Carson
Edmund Chaney
Richard Chapman
James R. Clopton
Gep Colletti
Paul T. Costa
Cynthia Cowgill
Daniel J. Cox
Rue Cromwell
Herb J. Cross
James Patrick Curran

Richard J. Davidson
Douglas A. Davis

John Lattimer Delk
Florence L. Denmark
Douglas R. Denney
Zachary Dershowitz
A. Devries
Robert R. Dies
Daniel Doleys
Edward Dougherty
Ralph M. Dreger
Thomas J. D'Zurilla

Craig Edelbrock
Jeff Edelson
Robert V. Erikson
Peter J. Equi
Stephen Etkind
Hans J. Eysenck

Norman Farberow
Amerigo Farina
A. David Feinstein
Robert D. Felner
C. B. Ferster
A. J. Finch, Jr.
Eva Fogelman
Donald C. Fowles
Susan Frank
Alfred S. Friedman
Ronald J. Friedman

Lawrence Gaines
John P. Galassi
Merna Dee Galassi
B. L. Gant
James B. Garrett
Robert J. Gatchel
Margaret J. Gatz
William Gayton
Joanne Gersten
Alan Gessner
Lewis R. Goldberg
Leo Goldberger
Marvin R. Goldfried
Gerald Goldstone
Irving Gottesman
John Gottman
John R. Graham
Wilson Guertin
Ruben C. Gur
Alan S. Gurman
Malcolm Gynther

Richard Haier
Judith Hall
Scott B. Hamilton
Constance Hanf

A. Hakstian
Donald P. Hartmann
Robert Hauser
Alfred Heilbrun
David E. Herr
Michel Hersen
Dorothy Hochreich
Martin L. Hoffman
Robert Hogan
Steven D. Holon
Philip Holzman
Hyman Hopps
John J. Horan
Kenneth Howard

Allen C. Israel
Robert J. Ivnik

Neil Jacobson
D. Balfour Jeffrey
Victor C. Joe
Warren Jones

Gershen Kaufman
Thomas D. Kennedy
Kevin Kennelly
Michael Keown
Bill N. Kinder
Irving Kirsch
Benjamin Kleinmuntz
Irving J. Knopf
Martin Kohn
Vladimir J. Konečni
Gerald P. Koocher
Elizabeth M. Koppitz
William M. Kurtines

Michael Lambert
Nadine Lambert
Harry A. Lando
Elaine L. LaMonica
Ellen Langer
Richard Lazarus
Herbert Lefcourt
Ellen Lenney
Gloria R. Leon
Mark R. Lepper
Hannah Lerman
Harry S. Levin
Eugene S. Levitt
Leonard S. Levitz
Howard Levanthal
Edward Lichtenstein
Richard Lippa
Leonard Lipton
Joseph Lopiccolo

Theodore W. Lorei
Raymond P. Lorian
Maurice Lorr
Bernard Lubin
Lester Luborsky

Alfred P. MacDonald, Jr.
Marion L. MacDonald
Stephen A. Maisto
Martin Manosevitz
K. Gerald Marsden
Barclay Martin
Raymond Martorano
Suzanne Martorano
Charles G. Matthews
Janet R. Matthews
Ralph Maurer
Alfred McAlister
Andrew McGhie
William McIntosh
Kate McMahon
Quinn McNemar
Paul McReynolds
William T. McReynolds
Hayden L. Mees
Donald H. Meichenbaum
Joseph Mendels
Lovick Miller
Peter M. Miller
William R. Miller
Iaian Montgomery
Richard Morris
Don Mosher
Thomas Mulholland
Louise Musser

Vincent J. Nerviano
Charles S. Newmark
Mike Nietzel
Stephen Nowicki
Vincent Nowlis
Jum C. Nunnally

Margaret O'Conner
Robert O'Conner
David E. Orienski
Jacob Lee Orlofsky
Gennaro Ottomanelli
John E. Overall

Susan Pepper
Thomas P. Petzel
Roy W. Persons
Douglas Powell

Janet Rafferty
Steven Reiss
Alexander Rich
C. Steven Richards
Frank C. Richardson
Manuel Riklan
David C. Rimm
Arnold Rincover
Ross Rizley
Sherry Rochester
Joanna Rohrbaugh
David Rosenthal
Alan Rosenwald
Timothy T. Ryan

Ann Salomon
Ina Samuels
Irwin Sandler
Victor D. Sanua
Irwin G. Sarason
Jerome Sattler
Earl Schaefer
Harold M. Schroder
Harold E. Schroeder
Lee Sechrest
M. E. P. Seligman
Lewis J. Sherman
Marianne L. Simmel
Arthur B. Silverstein

Jacob O. Sines
Harvey A. Skinner
Jonathan C. Smith
C. R. Snyder
Linda C. Sobell
John F. Solin
Paul A. Spiers
John J. Spinetta
Bonnie Spring
Fred F. Stauss
Abigail Stewart
Mark Stewart
Judith M. Stillion
Milton Strauss
Kenneth R. Suckerman
Norman D. Sundberg
Stanley Sue

Shelley E. Taylor
Francis Terrell
Robert M. Thorndike
John Todd
Craig T. Twentyman
Forrest Tyler

John Vincent

James A. Wakefield
Barbara Wallston
Charles G. Watson
John G. Way
David Wechsler
Richard Weigel
Carol K. Whalen
Cathy Spatz Widom
Jerry S. Wiggins
Wallace W. Wilkins

Antoinette M. Zeiss
Marvin Zuckerman
Miron Zuckerman

APA Journal Staff

Anita DeVivo
Executive Editor

Ann I. Mahoney
Manager, Journal Production

Barbara R. Richman
Production Supervisor

Robert J. Hayward
Advertising Representative

Anne Redman
Subscription Manager

Author Index to Volume 46

Key to Pagination

<i>Issue</i>	<i>Month</i>	<i>Pages</i>	<i>Issue</i>	<i>Month</i>	<i>Pages</i>
1	February	1-206	4	August	595-838
2	April	207-384	5	October	839-1169
3	June	385-594	6	December	1171-1591

ARTICLES

- Achenbach, Thomas M. The Child Behavior Profile: I. Boys aged 6-11..... 478
- Achenbach, Thomas M. Psychopathology of childhood: Research problems and issues..... 759
- Achterberg, Gloria. *See* Stedman, James M.
- Acosta, Frank X., Yamamoto, Joe, and Wilcox, Stuart A. Application of electromyographic biofeedback to the relaxation training of schizophrenic, neurotic, and tension headache patients..... 383
- Adams, Henry E. *See* Sturgis, Ellie T.
- Adams, John E. *See* Renson, Gisele J.
- Akamatsu, T. John, and Farudi, Parvis A. Effects of model status and juvenile offender type on the imitation of self-reward criteria..... 187
- Alcorn, John D. *See* Counts, D. Kenneth.
- Alden, Lynn. Factor analysis of the Ward Atmosphere Scale..... 175
- Alexander, Richard S. *See* Lachar, David
- Allain, A. *See* Sutker, P.
- Allain, Albert N. *See* Sutker, Patricia B.
- Allen, Deborah R. *See* Ferguson, Lucy Rau.
- Allen, George J. *See* Slutsky, Jeffrey M.
- Allgeier, Elizabeth Rice, and Fogel, Arthur F. Coital position and sex roles: Responses to cross-sex behavior in bed..... 588
- Alpert, Murray. *See* Anderson, Lowell.
- Amante, Dominic. Response to psychometric phrenology revisited: Comments on neuro-psychological testing..... 1491
- Anderson, Lowell, Dancis, Joseph, and Alpert, Murray. Behavioral contingencies and self-mutilation in Lesch-Nyhan disease..... 529
- Andrasik, Frank. *See* Holroyd, Kenneth A.
- Annis, Helen M. *See* Kohn, Paul M.
- Antista, Barbara J. *See* Kremsdorf, Ross B.
- Aranalde, Maria de los Angeles. *See* Szapocznik, José.
- Archer, R. *See* Sutker, P.
- Arkowitz, Hal. *See* Royce, W. Stephen.
- Ascher, L. Michael, and Efran, Jay S. Use of paradoxical intention in a behavioral program for sleep onset insomnia..... 547
- Atkins, Alvin. *See* Geller, Andrew M.
- Aved, Barbara M. *See* Weiss, Robert L.
- Bailey, Kent G. *See* Davidson, Kay M.
- Bailey, Kent G. Psychotherapy or massage parlor technology? Comments on the Zeiss, Rosen, and Zeiss treatment procedure..... 1502
- Ball, Leah R. *See* Kurtines, William M.
- Barbarin, Oscar A. Comparison of symbolic and overt aversion in the self-control of smoking. 1569
- Bardwick, Judith M. *See* O'Connor, Kevin.
- Barnes, Daniel F., and Berzins, Juris I. A and B undergraduate interviewers of schizophrenic and neurotic inpatients: A test of the interaction hypothesis..... 1368
- Barrett, Carol J. Effectiveness of widows' groups in facilitating change..... 20
- Bassett, John E. *See* Gayton, William F.
- Beach, Lee Roy. *See* Fiedler, Decky.
- Beck, Aaron T. *See* Bedrosian, Richard C.
- Bedrosian, Richard C., and Beck, Aaron T. Premature conclusions regarding black and white suicide attempters: A reply to Steele..... 1498
- Beiman, Irving, Israel, Eileen, and Johnson, Stephen A. During training and posttraining effects of live and taped extended progressive relaxation, self-relaxation, and electromyogram biofeedback..... 314
- Bentler, P. M. *See* Zukow, Patricia Goldring.

- Bentler, P. M., and Newcomb, Michael D. Longitudinal study of marital success and failure. . . 1053
- Berry, G. James. *See* Braucht, G. Nicholas.
- Berzins, Juris I. *See* Barnes, Daniel F.
- Berzins, Juris I., Welling, Martha A., and Wetter, Robert E. A new measure of psychological androgyny based on the Personality Research Form. . . 126
- Best, Helen. *See* Craig, Kenneth D.
- Best, J. Allan. *See* Craig, Kenneth D.
- Beutler, Larry E., Pollack, Stephen, and Jobe, Avis. "Acceptance," values, and therapeutic change. . . 198
- Bhatia, Kiran, and Golin, Sanford. Role of locus of control in frustration-produced aggression. . . 364
- Bigelow, Llewellyn B. *See* Donnelly, Edward F.
- Biglan, Anthony. *See* Lewis, Clifford E.
- Birkimer, John C. *See* Nelson, Wilbur J., Jr.
- Borkovec, T. D., Grayson, J. B., and Cooper, K. M. Treatment of general tension: Subjective and physiological effects of progressive relaxation. . . 518
- Braucht, G. Nicholas, Kirby, Michael W., and Berry, G. James. Psychosocial correlates of empirical types of multiple drug abusers. . . 1463
- Breen, Lawrence J., Endler, Norman S., Prociuk, Terry J., and Okada, Marilyn. Person \times Situation interaction in personality prediction: Some specifics of the person factor. . . 567
- Bromet, Evelyn. *See* Moos, Rudolf H.
- Brown, Gregory. *See* Wilson, Robert S.
- Brown, Robert A. *See* Kirshner, Barry J.
- Brownell, Kelly D., Heckerman, Carol L., and Westlake, Robert J. Therapist and group contact as variables in the behavioral treatment of obesity. . . 593
- Buck, Jeffrey A., and Graham, John R. The 4-3 MMPI profile type: A failure to replicate. . . 344
- Bumberry, William, Oliver, J. M., and McClure, James N. Validation of the Beck Depression Inventory in a university population using psychiatric estimate as the criterion. . . 150
- Burkhart, Barry R. *See* Christian, William L.
- Burkhart, Barry R. *See* Schwarz, Raymond M.
- Burkhart, Barry R., Schwarz, Raymond M., and Green, Samuel B. Relationships between dimensions of anxiety and sensation seeking. . . 194
- Burns, Joan W. *See* King, Glen D.
- Butcher, James N., and Tellegen, Auke. Common methodological problems in MMPI research. . . 620
- Butt, James H. *See* Shipley, Robert H.
- Caffey, Eugene M., Jr. *See* Lorei, Theodore W.
- Campbell, Susan B., and Steinert, Yvonne. Comparisons of rating scales of child psychopathology in clinic and nonclinic samples. . . 358
- Cardwell, Gilbert F. *See* Greenberg, Roger P.
- Carpenter, Paul J. *See* Del Gaudio, Andrew C.
- Carroll, Charles F., and Reppucci, N. Dickon. Meanings that professionals attach to labels for children. . . 372
- Carroll, Jerome F. X., Klein, M. Israel, and Santo, Yoav. Comparison of the similarities and differences in the self-concepts of male alcoholics and addicts. . . 575
- Chamberlain, C. J. *See* Joslyn, Dan.
- Chaney, Edmund F., O'Leary, Michael R., and Marlatt, G. Alan. Skill training with alcoholics. . . 1092
- Chapman, Jean P. *See* Hertler, Chris A.
- Chapman, Loren J. *See* Hertler, Chris A.
- Chapman, Stanley L., and Jeffrey, D. Balfour. Situational management, standard setting, and self-reward in a behavior modification weight loss program. . . 1588
- Charles, H. L. *See* Penk, W. E.
- Chernovetz, Mary Ellen O'C. *See* Jones, Warren H.
- Chirico, Bernard M. *See* Kendall, Philip C.
- Christ, Maryann. *See* Stalonas, Peter M., Jr.
- Christian, Kenneth W. Aspects of the self-concept related to level of self-esteem. . . 1151
- Christian, William L., Burkhart, Barry R., and Gynther, Malcolm D. Subtle-obvious ratings of MMPI items: New interest in an old concept. . . 1178
- Cicala, George A. *See* Mellstrom, Martin Jr.
- Ciminero, Anthony R. *See* Williams, Carolyn L.
- Clingempeel, W. Glenn. *See* Reppucci, N. Dickon.
- Clopton, James R. A note on the MMPI as a suicide predictor. . . 335
- Clopton, James R., and Klein, Gary L. An initial look at the redundancy of specialized MMPI scales. . . 1436
- Clum, George A. Personality traits and environmental variables as independent predictors of posthospitalization outcome. . . 839
- Cole, David, Rodriguez, Jacqueline, and Cole, Shirley. Locus of control in Mexicans and Chicanos: The case of the missing fatalist. . . 1323
- Cole, Shirley. *See* Cole, David.
- Condiotte, Mark. *See* Leary, John.
- Cohen, Gary H. *See* Sutker, Patricia B.
- Comrey, Andrew L. Common methodological problems in factor analytic studies. . . 648

- Cooper, K. M. *See* Borkovec, T. D.
- Cortner, Robert H. *See* Stedman, James M.
- Costello, Raymond M. Premorbid social competence construct generalizability across ethnic groups: Path analyses with two premorbid social competence components... 1164
- Counts, D. Kenneth, Hollandsworth, James G., Jr., and Alcorn, John D. Use of electromyographic biofeedback and cue-controlled relaxation in the treatment of test anxiety... 990
- Cowen, Emory L. Some problems in community program evaluation research... 792
- Cowan, Philip A. *See* Dyckman, John M.
- Cox, W. Miles. Frequent citations in the *Journal of Consulting and Clinical Psychology* during the 1970s... 204
- Craig, Kenneth D. *See* Nemetz, Georgia H.
- Craig, Kenneth D., Best, Helen, and Best, J. Allan. Self-regulatory effects of monitoring sensory and affective dimensions of pain... 573
- Cronkite, Ruth C., and Moos, Rudolf H. Evaluating alcoholism treatment programs: An integrated approach... 1105
- Dahlberg, Charles. *See* Natale, Michael.
- Dahlstrom, W. Grant. *See* Gynther, Malcolm D.
- Dancis, Joseph. *See* Anderson, Lowell.
- Davidson, Kay M., and Bailey, Kent G. Effects of "status sets" on Rotter's Locus of Control scale... 186
- Davidson, Gerald C. Not can but ought: The treatment of homosexuality... 170
- Dean, Raymond S. Distinguishing learning-disabled and emotionally disturbed children on the WISC-R... 381
- DeHorn, Allan, and Klinge, Valerie. Correlations and factor analysis of the WISC-R and the Peabody Picture Vocabulary Test for an adolescent psychiatric sample... 1160
- Del Gaudio, Andrew C., Carpenter, Paul J., and Morrow, Gary R. Male and female treatment differences: Can they be generalized?... 1577
- Denney, Douglas R. *See* Elliott, Charles H.
- Dies, Robert R. *See* Kirchner, Barry J.
- Dikmen, Sureyya S. *See* Dodrill, Carl B.
- Dodrill, Carl B. The hand dynamometer as a neuropsychological measure... 1432
- Dodrill, Carl B., and Dikmen, Sureyya S. The Seashore Tonal Memory Test as a neuropsychological measure... 192
- Donahoe, Clyde P., Jr. *See* Freedman, Barbara J.
- Donnelly, Edward F., Nasrallah, Henry A., Wyatt, Richard J., Gullin, J. Christian, and Bigelow, Llewellyn B. Effects of dopamine synthesis inhibition on WAIS comprehension... 385
- Dor-Shav, Netta Kohn. On the long-range effects of concentration camp internment on Nazi victims: 25 years later... 1
- Douglass, Frazier M., IV. *See* Khavari, Khalil A.
- Dudley, Gary E. Effects of sex, social desirability, and birth order on the Defense Mechanisms Inventory... 1419
- Dunivant, Noel. *See* Sadd, Susan.
- Dyckman, John M., and Cowan, Philip A. Imaging vividness and the outcome of in vivo and imagined scene desensitization... 1155
- Early, Charles E., and Kleinknecht, Ronald A. The Palmar Sweat Index as a function of repression-sensitization and fear of dentistry... 184
- Eberly, Carole. *See* Kendall, Philip C.
- Eddy, Sara. *See* Rierdan, Jill.
- Edinger, Jack. *See* Kendall, Philip C.
- Edwards, Keith J. *See* Roe, John E.
- Efran, Jay S. *See* Ascher, L. Michael.
- Eisler, Richard M. *See* Pachman, Joseph S.
- Elliott, Charles H., and Denney, Douglas R. A multiple-component treatment approach to smoking reduction... 1330
- Endelman, Janet R. *See* Snyder, C. R.
- Endler, Norman S. *See* Breen, Lawrence J.
- Epstein, Norman, and Jackson, Elizabeth. An outcome study of short-term communication training with married couples... 207
- Epstein, Seymour. Avoidance-approach: The fifth basic conflict... 1016
- Eron, Leonard D. *See* Huesmann, L. Rowell.
- Ease, John T., and Wilkins, Wallace. Empathy and imagery in avoidance behavior reduction... 202
- Eysenck, H. J. *See* Zuckerman, Marvin.
- Eysenck, Sybil. *See* Zuckerman, Marvin.
- Farbry, John E. *See* Shipley, Robert H.
- Farina, Amerigo, Murray, Pauline J., and Groh, Thomas. Sex and worker acceptance of a former mental patient... 887
- Farkas, Gary M. Comments on Levin et al. and Rosen and Kopel: Internal and external validity issues... 1515
- Farudi, Parvis A. *See* Akamatsu, T. John.

- Ferguson, Lucy Rau, and Allen, Deborah R. Congruence of parental perception, marital satisfaction, and child adjustment..... 345
- Feyerabend, C. *See* Russell, M. A. H.
- Fibell, Bobbi, and Hale, W. Daniel. The generalized Expectancy for Success Scale—a new measure..... 924
- Fiedler, Decky, and Beach, Lee Roy. On the decision to be assertive..... 537
- Finch, A. J., Jr. *See* Kendall, Philip C.
- Finch, Alfred J., Jr. *See* Newmark, Charles S.
- Fioravanti, Mario. *See* Gough, Harrison G.
- Fleece, L. *See* Melamed, B. G.
- Fogel, Arthur F. *See* Allgeier, Elizabeth Rice.
- Folkins, Carlyle H. *See* Miller, Stephen H.
- Follingstad, Diane R. *See* Moreault, Denise.
- Ford, Julian D. Therapeutic relationship in behavior therapy: An empirical analysis..... 1302
- Forehand, Rex, Wells, Karen C., and Sturgis, Ellie T. Predictors of child noncompliant behavior in the home..... 179
- Forsythe, Alan B. *See* Tuma, A. Hussain.
- Foy, David W. *See* Pachman, Joseph S.
- Freedman, Barbara J., Rosenthal, Lisa, Donahoe, Clyde P., Jr., Schlundt, David G., and McFall, Richard M. A social-behavioral analysis of skill deficits in delinquent and nondelinquent adolescent boys..... 1448
- Furman Wyndol. *See* Kelly, Jeffrey A.
- Gambara, Salvatore. *See* Levin, Saul M.
- Garfield, Sol L. *See* Kurtz, Richard M.
- Garfield, Sol L. Research problems in clinical diagnosis..... 596
- Garrity, Linda I., Servos, Andria B. Comparison of measures of adaptive behaviors in preschool children..... 288
- Garske, John P. *See* Kowitt, Michael R.
- Gayton, William F., Bassett, John E., Tavormina, Joseph, and Ozmon, Kenneth L. Repression-sensitization and health behavior..... 1542
- Geller, Andrew M., and Atkins, Alvin. Cognitive and personality factors in suicidal behavior.. 860
- Geller, Jesse D. *See* Goodwin, William B.
- Gelman, Richard, and McGinley, Hugh. Interpersonal liking and self-disclosure..... 1549
- Geyer, Scott. *See* Sutker, Patricia B.
- Giannetti, Ronald A., Johnson, James H., Klingler, Daniel E., and Williams, Thomas A. Comparison of linear and configural MMPI diagnostic methods with an uncontaminated criterion 1046
- Gilbert, Gary S. *See* Parker, Jerry C.
- Gilllin, J. Christian. *See* Donnelly, Edward F.
- Gilmer, Timothy M., and Reid, David W. Locus of control, prediction, and performance on university examinations..... 565
- Girodo, Michel, and Roehl, Julius. Cognitive preparation and coping self-talk: Anxiety management during the stress of flying..... 978
- Glasgow, Russell E. Effects of a self-control manual, rapid smoking, and amount of therapist contact on smoking reduction..... 1439
- Glennon, Blair, and Weisz, John R. An observational approach to the assessment of anxiety in young children..... 1246
- Golden, Charles J. *See* Purisch, Arnold D.
- Golden, Charles J., Hammeke, Thomas A., and Purisch, Arnold D. Diagnostic validity of a standardized neuropsychological battery derived from Luria's neuropsychological tests..... 1258
- Goldfried, Marvin R., Linehan, Marsha M., and Smith, Jean L. Reduction of test anxiety through cognitive restructuring..... 32
- Golin, Sanford. *See* Bhatia, Kiran.
- Gomes-Schwartz, Beverly. Effective ingredients in psychotherapy: Prediction of outcome from process variables..... 1023
- Gomes-Schwartz, Beverly, and Schwartz, Joseph M. Psychotherapy process variables distinguishing the "inherently helpful" person from the professional psychotherapist..... 196
- Good, Michael I. *See* Surwit, Richard S.
- Goodfriend, Shirley. *See* Horowitz, Leonard M.
- Goodwin, William B. Psychotherapeutic styles: Self-perceptions of therapists varying on A-B type and experience..... 1590
- Goodwin, William B., Quinlan, Donald M., and Geller, Jesse D. Practicing A and B psychotherapists' responses to schizoid and neurotic patient prototypes..... 356
- Gordon, Robert M. Imprecision or dissonance? A reply to Harris and Harvey..... 329
- Gottman, John M. *See* Levenson, Robert W.
- Gough, Harrison G., Lazzari, Renato, and Fioravanti, Mario. Self versus ideal self: A comparison of five Adjective Check List indices..... 1085
- Graham, John R. *See* Buck, Jeffrey A.
- Grayson, J. B. *See* Borkovec, T. D.
- Green, Samuel B. *See* Burkhart, Barry R.
- Green, Samuel B. *See* Schwarz, Raymond M.

AUTHOR INDEX TO VOLUME 46

Greenberg, Roger P., and Cardwell, Gilbert F. Rorschach developmental level and intelligence factors.	844
Greene, Roger L. Can clients provide valuable feedback to clinicians about their personality interpretation? Greene replies.	1496
Grimm, Laurence G. See Kanfer, Frederick H.	
Grisell, James. See Wilson, Robert S.	
Groh, Thomas. See Farina, Amerigo.	
Grundvig, John L. See Joslyn, Dan.	
Gutkin, Terry B. Some useful statistics for the interpretation of the WISC-R	1561
Gynther, Malcolm D. See Christian, William L.	
Gynther, Malcolm D., Lachar, David, and Dahlstrom, W. Grant. Are special norms for minorities needed? Development of an MMPI F scale for blacks	1403
Hale, W. Daniel. See Fibel, Bobbi.	
Hammeke, Thomas A. See Golden, Charles J.	
Hammeke, Thomas A. See Purisch, Arnold D.	
Handelsman, Mitchell M. See Snyder, C. R.	
Hannay, H. Julia. See King, Glen D.	
Hansson, Robert O. See Jones, Warren H.	
Harford, Robert J. Drug preferences of multiple drug abusers	908
Harris, Ben, and Harvey, John H. Social psychological concepts applied to clinical processes: On the need for precision.	376
Hart, Russell R. Therapeutic effectiveness of setting and monitoring goals.	1242
Harvey, John H. See Harris, Ben.	
Hauri, Peter. See Silberfarb, Peter M.	
Hawes, R. See Melamed, B. G.	
Heaton, Robert K., Smith, Harold H., Jr., Lehman, Ralph A. W., and Vogt, Arthur T. Prospects for faking believable deficits on neuropsychological testing.	892
Heckerman, Carol L. See Brownell, Kelly D.	
Heilbrun, Alfred B., Jr. Projective and repressive styles of processing aversive information	156
Hertler, Chris A., Chapman, Loren J., and Chapman, Jean P. A scoring manual for literalness in proverb interpretation.	551
Hile, Matthew G. See Nietzel, Michael T.	
Hiscock, Merrill. Imagery assessment through self-report: What do imagery questionnaires measure?	273
Hobfoll, Stevan E., and Penner, Louis A. Effect of physical attractiveness on therapists' initial judgments of a person's self-concept.	177
Hochreich, Dorothy J. Defensive externality and level of aspiration	1187
Hoffmann, Norman G. See Overall, John E.	
Hoffmann, Norman G., and Overall, Peggy B. Factor structure of the SCL-90 in a psychiatric clinic population.	577
Holland, Terrill R. Dimensions, patterns, and personality correlates of drug abuse in an offender population.	1036
Hollandsworth, James G., Jr. See Counts, D. Kenneth.	
Holroyd, Kenneth A., and Andrasik, Frank. Coping and the self-control of chronic tension headache.	356
Holzmueller, Ana. See Wiggins, Jerry S.	
Horowitz, Leonard M., Sampson, Harold, Siegelman, Ellen Y., Weiss, Joseph, and Goodfriend, Shirley. Cohesive and dispersal behaviors: Two classes of concomitant change in psychotherapy.	1071
Horwitz, Bruce. See Shipley, Robert H.	
Hosch, Harmon M. See Trevithick, Lee.	
Huesmann, L. Rowell, Lefkowitz, Monroe M., and Eron, Leonard D. Sum of MMPI Scales P, 4, and 9 as a measure of aggression.	442
Hutcherson, S. See Melamed, B. G.	
Intagliata, James C. Increasing the interpersonal problem-solving skills of an alcoholic population.	913
Israel, Eileen. See Beiman, Irving.	
Ivnik, Robert J. Neuropsychological stability in multiple sclerosis	368
Jackson, Douglas N. See Skinner, Harvey A.	
Jackson, Elizabeth. See Epstein, Norman.	
Jacobson, Neil S. Specific and nonspecific factors in the effectiveness of a behavioral approach to the treatment of marital discord.	
Jaffe, Joseph. See Natale, Michael.	
Janda, Lois. See O'Grady, Kevin.	
Jeffrey, D. Balfour. See Chapman, Stanley L.	
Jeffery, Robert W., Vender, Michael, Wing, Rena R. Weight loss and behavior change 1 year after behavioral treatment for obesity.	
Jew, Charles. See Shawver, Lois.	
Jobe, Avis. See Bentler, Larry E.	

AUTHOR INDEX TO VOLUME 46

Johnson, James H.	See Giannetti, Ronald A.	
Johnson, James H.	See Sarason, Irwin G.	
Johnson, James H.	See Smith, Ronald E.	
Johnson, Stephen A.	See Beiman, Irving.	
Johnson, William G.	See Stalonas, Peter M., Jr.	
Jones, Warren H., Chernovetz, Mary Ellen O'C., and Hansson, Robert O.	The enigma of androgyny: Differential implications for males and females?	298
Joslyn, Dan, Grundvig, John L., and Chamberlain, C. J.	Predicting confabulation from the Graham-Kendall Memory-For-Designs Test	181
Kanfer, Frederick H., and Grimm, Laurence G.	Freedom of choice and behavioral change	873
Kantorowitz, David A., Walters, Joyce, and Pezdek, Kathy.	Positive versus negative self-monitoring in the self-control of smoking	1148
Kay, Edwin J., Lyons, Arthur, Newman, William, Mankin, Donald, and Loeb, Roger C.	A longitudinal study of the personality correlates of marijuana use	470
Kazdin, Alan E.	Evaluating the generality of findings in analogue therapy research	673
Kazdin, Alan E.	Methodological and interpretative problems of single-case experimental designs	629
Keeley, Stuart M.	See Waller, Ronald W.	
Kellerman, Jonathan.	A note on psychosomatic factors in the etiology of neoplasms	1522
Kelly, Jeffrey A., Furman, Wyndol, and Young, Veronica.	Problems associated with the typological measurement of sex roles and androgyny	1574
Kendall, Philip C.	See Newmark, Charles S.	
Kendall, Philip C.	Anxiety: States, traits—Situations?	280
Kendall, Philip C., Edinger, Jack, and Eberly, Carole.	Taylor's MMPI correction factor for spinal cord injury: Empirical endorsement	370
Kendall, Philip C., Finch, A. J., Jr., Little, Verda L., Chirico, Bernard M., and Ollendick, Thomas H.	Variations in a construct: Quantitative and qualitative differences in children's locus of control	590
Kendall, Philip C., and Finch, A. J., Jr.	A cognitive-behavioral treatment for impulsivity: A group comparison study	110
Kendall, Philip C., Finch, A. J., Jr., and Montgomery, L. E.	Vicarious anxiety: A systematic evaluation of a vicarious threat to self-esteem	997
Keyson, Mae.	See Turner, Robert G.	
Khavari, Khalil A., and Douglass, Frazier M., IV.	The Polydrug Assessment Scale: A psychometric technique for the indirect measurement of drug use	1566
King, Glen D., Hannay, H. Julia, Masek, Bruce J., and Burns, Joan W.	Effects of anxiety and sex on neuropsychological tests	375
Kirby, Michael W.	See Braucht, G. Nicholas.	
Kirshner, Barry J., Dies, Robert R., and Brown, Robert A.	Effects of experimental manipulation of self-disclosure on group cohesiveness	1171
Klein, Gary L.	See Clopton, James R.	
Klein, M. Israel.	See Carroll, Jerome F. X.	
Kleinknecht, Ronald A.	See Early, Charles E.	
Klinge, Valerie.	See DeHorn, Allan.	
Klingler, Daniel E.	See Giannetti, Ronald A.	
Knight, Raymond.	See Roff, James D.	
Kohn, Paul M., and Annis, Helen M.	Personality and social factors in adolescent marijuana use: A path-analytic study	366
Kondo, Charles Y.	See Nietzel, Michael T.	
Kopel, Steven A.	See Rosen, Raymond C.	
Kowitt, Michael.	See Natale, Michael.	
Kowitt, Michael R., and Garske, John P.	Modality, self-disclosure, and gender as determinants of psychotherapeutic attraction	173
Kremsdorff, Ross B., Palladino, Lucy J., Polenz, Douglas D., and Antista, Barbara J.	Effects of the sex of both interviewer and subject on reported manifest dream content	1166
Kurtines, William.	See Szapocznik, José.	
Kurtines, William M., Ball, Leah R., and Wood, Gloria H.	Personality characteristics of long-term recovered alcoholics: A comparative analysis	971
Kurtz, Richard M., and Garfield, Sol L.	Illusory correlation: A further exploration of Chapman's paradigm	1009
Kushner, Kenneth.	On the external validity of two psychotherapy analogues	1394
Lachar, David.	See Gynther, Malcolm D.	
Lachar, David, and Alexander, Richard, S.	Veridicality of self-report: Replicated correlates of the Wiggins MMPI content scales	1349
Lando, Harry A., and McCullough, John A.	Clinical application of a broad-spectrum behavioral approach to chronic smokers	1583
Lang, Elizabeth.	See Rierdan, Jill.	
Lang, Peter J.	See Weerts, Theodore C.	
Lang, Rudie J.	Multivariate classification of day-care patients: Personality as a dimensional continuum	1212
Lansky, David.	See Nathan, Peter E.	

- Lansky, David, Nathan, Peter E., and Lawson, David M. Blood alcohol level discrimination by alcoholics: The role of internal and external cues. 953
- Lawlis, G. Frank. See Stedman, James M.
- Lawson, David M. See Lansky, David.
- Lazzari, Renato. See Gough, Harrison G.
- Lefkowitz, Monroe M. See Huesmann, L. Rowell.
- Lehman, Ralph A. W. See Heaton, Robert K.
- Lehman, Robert E. Symptom contamination of the Schedule of Recent Events. 1564
- Lehrer, Paul M. Psychophysiological effects of progressive relaxation in anxiety neurotic patients and of progressive relaxation and alpha feedback in nonpatients. 389
- Lenauer, Michael. See Sadd, Susan.
- Leonard, Calista V. Response to "A note on the MMPI as a suicide predictor". 337
- Levenson, Robert W., and Gottman, John M. Toward the assessment of social competence. 453
- Levin, Harvey. See Overall, John E.
- Levin, Saul M., Gambaro, Salvatore, and Wolfensohn, Lawrence. Penile tumescence as a measure of sexual arousal: A reply to Farkas. 1517
- Lewis, Clifford E., Biglan, Anthony, and Steinbock, Elizabeth. Self-administered relaxation training and money deposits in the treatment of recurrent anxiety. 1274
- Licht, Mark H. See Paul, Gordon L.
- Lick, John, Condiotte, Mark, and Unger, Thomas. Effects of uncertainty about the behavior of a phobic stimulus on subjects' fear reactions. 1559
- Linehan, Marsha M. See Goldfried, Marvin R.
- Little, Verda L. See Kendall, Philip C.
- Loeb, Roger C. See Kay, Edwin J.
- Longabaugh, Richard. See Vannicelli, Marsha.
- Lorei, Theodore W., and Caffey, Eugene M., Jr. Goal definition by staff consensus: A contribution to the planning, delivery, and evaluation of mental health services. 1284
- Lubin, Bernard, Marone, Joseph G., and Nathan, Ronald G. Comparison of self-administered and examiner-administered Depression Adjective Check Lists. 584
- Lyle, J. G. See Rawling, P.
- Lyons, Arthur. See Kay, Edwin J.
- Maher, Brendan A. Preface. 395
- Maher, Brendan A. A reader's, writer's, and reviewer's guide to assessing research reports in clinical psychology. 835
- Maher, Brendan A. Stimulus sampling in clinical research: Representative design reviewed. 643
- Mahoney, Michael J. Experimental methods and outcome evaluation. 660
- Maloney, Lorrie J. See Spinetta, John J.
- Mankin, Donald. See Kay, Edwin J.
- Mann, David W. See O'Connor, Kevin.
- Margolin, Gayla. Relationships among marital assessment procedures: A correlational study. 1556
- Margolin, Gayla, and Weiss, Robert L. Comparative evaluation of therapeutic components associated with behavioral treatments. 1476
- Marlatt, G. Alan. See Chaney, Edmund F.
- Marone, Joseph G. See Lubin, Bernard.
- Martin, April. See Silverman, Lloyd H.
- Martindale, Colin. Hemispheric asymmetry and Jewish intelligence patterns. 1299
- Martindale, Colin. The therapist-as-fixed-effect fallacy in psychotherapy research. 1526
- Masek, Bruce J. See King, Glen D.
- Massey, Frank. See Pachman, Joseph S.
- Matarazzo, Joseph D., Wiens, Arthur N., and Shealy, Allen E. Correlation of WAIS IQ in 10 pairs of brothers. 571
- May, Philip R. A. See Tuma, A. Hussain.
- McClure, James N. See Bumberry, William.
- McCreary, Charles. See Turner, Judith.
- McCullough, John A. See Lando, Harry A.
- McDowell, David J. Bias or artifact? A reply to Reade and Wertheimer. 1487
- McFall, Richard M. See Freedman, Barbara J.
- McFall, Richard M. Smoking-cessation research. 703
- McGinley, Hugh. See Gelman, Richard.
- McIntyre, Curtis W. See Shostak, David A.
- McSharry, Dennis M. See Simons, Lynn S.
- Meehl, Paul E. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. 806
- Melamed, B. G., Yurcheson, R., Fleece, L., Hutcherson, S., and Hawes, R. Effects of film modeling on the reduction of anxiety-related behaviors in individuals varying in level of previous experience in the stress situation. 1357
- Mellstrom, Martin, Jr., Zuckerman, Marvin, and Cicala, George A. General versus specific traits in the assessment of anxiety. 423
- Mendelsohn, Eric. See Silverman, Lloyd H.
- Miller, Glenn A., and Nuesse, William. Characteristics of the emotional responsiveness of sensitizers and repressors to social stimuli. 339

- Miller, John E. *See* Shiek, David A.
- Miller, Martha P., Murphy, Philip J., and Miller, Terry P. Comparison of electromyographic feedback and progressive relaxation training in treating circumscribed anxiety stress reactions 1291
- Miller, Stephen H., O'Reilly, Charles A., Roberts, Karlene H., and Folkins, Carlyle H. Factor structure and scale reliabilities of the Adjective Check List across time. 189
- Miller, Terry P. *See* Miller, Martha P.
- Miller, William R. Behavioral treatment of problem drinkers: A comparative outcome study of three controlled drinking therapies. 74
- Mishara, Brian L. Geriatric patients who improve in token economy and general milieu treatment programs: A multivariate analysis. 1340
- Montgomery, L. E. *See* Kendall, Philip C.
- Moos, Rudolf H. *See* Cronkite, Ruth C.
- Moos, Rudolf H., and Bromet, Evelyn. Relation of patient attributes to perceptions of the treatment environment. 350
- Moreault, Denise, and Follingstad, Diane R. Sexual fantasies of females as a function of sex guilt and experimental response cues. 1385
- Morrow, Gary R. *See* Del Gaudio, Andrew C.
- Morton, Teru L. *See* Simons, Lynn S.
- Murphy, Philip J. *See* Miller, Martha P.
- Murray, Pauline J. *See* Farina, Amerigo.
- Nasrallah, Henry A. *See* Donnelly, Edward F.
- Natale, Michael, Dahlberg, Charles C., and Jaffe, Joseph. Effect of psychotomimetics (LSD and dextroamphetamine) on the use of primary- and secondary-process language. 352
- Natale, Michael, Kowitt, Michael, Dahlberg, Charles, and Jaffe, Joseph. Effect of psychotomimetics (LSD and dextroamphetamine) on the use of figurative language during psychoanalysis. 1579
- Nathan, Peter E. *See* Lansky, David.
- Nathan, Peter E., and Lansky, David. Common methodological problems in research on the addictions. 713
- Nathan, Ronald G. *See* Lubin, Bernard.
- Neale, John M. *See* Stone, Arthur A.
- Nelson, Wilbur J., Jr., and Birkimer, John C. Role of self-instruction and self-reinforcement in the modification of impulsivity. 183
- Nemec, Richard E. Effects of controlled background interference on test performance by right and left hemiplegics. 294
- Nemetz, Georgia H., Craig, Kenneth D., and Reith, Gunther. Treatment of female sexual dysfunction through symbolic modeling. 62
- Newcomb, Michael D. *See* Bentler, P. M.
- Newman, William. *See* Kay, Edwin J.
- Newmark, Charles S., Ziff, David R., Finch, Alfred J., Jr., and Kendall, Philip C. Comparing the empirical validity of the standard form with two abbreviated MMPIs. 53
- Nietzel, Michael T., Hile, Matthew G., and Kondo, Charles Y. Diversity among lower-class therapy clients: A comparison of Class IV and Class V psychotherapy recipients. 377
- Nolan, J. Dennis, and Sandman, Curt. "Biosyntonic" therapy: Modification of an operant conditioning approach to pedophilia. 1133
- Nowicki, Stephen, Jr. Reported stressful events during developmental periods and their relation to locus of control orientation in college students. 1552
- Nuesse, William. *See* Miller, Glenn A.
- Obitz, Frederick W. Normative data on locus of control in male alcoholics undergoing treatment 379
- O'Connor, Kevin, Mann, David W., and Bardwick, Judith M. Androgyny and self-esteem in the upper middle class: A replication of Spence. 1168
- O'Grady, Kevin, and Janda, Louis. Psychometric correlates of the Mosher Forced Choice Guilt Inventory. 1581
- Okada, Marilyn. *See* Breen, Lawrence J.
- O'Leary, K. Daniel, and Turkewitz, Hillary. Methodological errors in marital and child treatment research. 747
- O'Leary, Michael R. *See* Chaney, Edmund F.
- Oliver, J. M. *See* Bumberry, William.
- Ollendick, Thomas H. *See* Kendall, Philip C.
- O'Reilly, Charles A. *See* Miller, Stephen H.
- Ottomanelli, Gennaro, Wilson, Peter, and Whyte, Richard. MMPI evaluation of 5-year methadone treatment status. 579
- Overall, John E., Hoffmann, Norman G., and Levin, Harvey. Effects of aging, organicity, alcoholism, and functional psychopathology on WAIS subtest profiles. 1315
- Overall, Peggy B. *See* Hoffmann, Norman G.
- Ozmon, Kenneth L. *See* Gayton, William F.
- Pachman, Joseph S., Foy, David W., Massey, Frank, and Eisler, Richard M. A factor analysis of assertive behaviors. 347
- Palladino, Lucy J. *See* Kremsdorf, Ross B.

- Parker, Jerry C., Gilbert, Gary S., and Thoreson, Richard W. Reduction of autonomic arousal in alcoholics: A comparison of relaxation and meditation techniques 879
- Parsons, Oscar A., and Prigatano, George P. Methodological considerations in clinical neuropsychological research. 608
- Paul, Gordon L., and Licht, Mark H. Resurrection of uniformity assumption myths and the fallacy of statistical absolutes in psychotherapy research. 1331
- Penk, W. E., Charles, H. L., and Van Hoose, T. A. Comparative effectiveness of day hospital and inpatient psychiatric treatment. 94
- Penner, Louis A. See Hobfoll, Stevan E.
- Pezdek, Kathy. See Kantorowitz, David A.
- Phelps, Patricia J. See Silberfarb, Peter M.
- Piotrowski, Richard J. Abnormality of subtest score differences on the WISC-R. 569
- Pleml, Duane. See Watson, Charles G.
- Polsnz, Douglas D. See Kremsdorf, Ross B.
- Poliack, Stephen. See Beutler, Larry E.
- Post, Amy L., Wittmaier, Bruce C., and Radin, Mitchell E. Self-disclosure as a function of state and trait anxiety. 12
- Poythress, Norman G., Jr. Selecting a short form of the MMPI: Addendum to Fachingbauer 331
- Prigatano, George P. See Parsons, Oscar A.
- Pritchard, David A. See Rosenblatt, Arthur I.
- Prociuk, Terry J. See Breen, Lawrence J.
- Purisch, Arnold D. See Golden, Charles J.
- Purisch, Arnold D., Golden, Charles J., and Hammeke, Thomas A. Discrimination of schizophrenic and brain-injured patients by a standardized version of Luria's neuropsychological tests. 1266
- Quinlan, Donald M. See Goodwin, William B.
- Radin, Mitchell E. See Post, Amy L.
- Raw, Martin. See Russell, M. A. H.
- Rawling, P., and Lyle, J. G. Cued recall and discrimination of memory deficit. 1227
- Rehm, Lynn P. Mood, pleasant events, and unpleasant events: Two pilot studies. 854
- Reid, David W. See Gilmor, Timothy M.
- Reith, Gunther. See Nemetz, Georgia H.
- Renson, Gisele J., Adams, John E., and Tinklenberg, Jared R. Buss-Durkee assessment and validation with violent versus nonviolent chronic alcohol abusers. 360
- Reppucci, N. Dickon, and Clingempeel, W. Glenn. Methodological issues in research with correctional populations. 727
- Reschly, Daniel J. WISC-R factor structures among Anglos, Blacks, Chicanos, and Native-American Papagos. 417
- Reppucci, N. Dickon. See Carroll, Charles F.
- Reynolds, William M. A question as to validity of Verbal Scale IQ as a WAIS short form 1535
- Rierdan, Jill, Lang, Elizabeth, and Eddy, Sara. Suicide and transparency responses on the Rorschach: A replication. 1162
- Roberts, Karlene H. See Miller, Stephen H.
- Rodriguez, Jacqueline. See Cole, David.
- Roe, John E., and Edwards, Keith J. The relationship of two process measurement systems for group therapy. 1545
- Roehl, Julius. See Girodo, Michel.
- Roff, James D., and Knight, Raymond. Young adult schizophrenics: Prediction of outcome and antecedent childhood factors. 947
- Rosen, Raymond C., and Kopel, Steven A. Role of penile tumescence measurement in the behavioral treatment of sexual deviation: Issues of validity. 1519
- Rosenbaum, Gerald. See Wilson, Robert S.
- Rosenblatt, Arthur I., and Pritchard, David A. Moderators of racial differences on the MMPI 1572
- Rosenthal, Lisa. See Freedman, Barbara J.
- Rourke, Daniel. See Wilson, Robert S.
- Royce, W. Stephen, and Arkowitz, Hal. Multimodal evaluation of practice interactions as treatment for social isolation. 239
- Russell, M. A. H., Raw, Martin, Taylor, C., Feyerabend, C., and Saloojee, Y. Blood nicotine and carboxyhaemoglobin levels after rapid-smoking aversion therapy 1423
- Sadd, Susan, Lenauer, Michael, Shaver, Phillip, and Dunivant, Noel. Objective measurement of fear of success and fear of failure: A factor analytic approach. 405
- Saloojee, Y. See Russell, M. A. H.
- Sampson, Harold. See Horowitz, Leonard M.
- Sandman, Curt. See Nolan, J. Dennis.
- Santo, Yoav. See Carroll, Jerome F. X.
- Sarason, Irwin G. See Smith, Ronald E.
- Sarason, Irwin G., Johnson, James H., and Siegel, Judith M. Assessing the impact of life changes: Development of the Life Experiences Survey. 932
- Sarason, Irwin G., and Stoops, Rick. Test anxiety and the passage of time. 102

- Scheff, Betty-Jane. *See* Vannicelli, Marsha.
- Schlundt, David G. *See* Freedman, Barbara J.
- Schuld, Donald. *See* Watson, Charles G.
- Schuldt, W. John. *See* Souheaver, Gary T.
- Schwartz, Joseph M. *See* Gomes-Schwartz, Beverly.
- Schwarz, J. Conrad. *See* Zuroff, David C.
- Schwarz, Raymond M. *See* Burkhart, Barry R.
- Schwarz, Raymond M., Burkhart, Barry R., and Green, Samuel B. Turning on or turning off: Sensation seeking or tension reduction as motivational determinants of alcohol use. 1144
- Scopetta, Mercedes Arca. *See* Szapocznik, José.
- Servos, Andria B. *See* Garrity, Linda I.
- Shapiro, David. *See* Surwit, Richard S.
- Shaver, Phillip. *See* Sadd, Susan.
- Shawver, Lois, and Jew, Charles. Predicting violent behavior from WAIS characteristics: A replication failure. 206
- Shealy, Allen E. *See* Matarazzo, Joseph D.
- Shiek, David A., and Miller, John E. Validity generalization of the WISC-R factor structure with 10½-year-old children. 583
- Shipley, Robert H., Butt, James H., Horwitz, Bruce, and Farby, John E. Preparation for a stressful medical procedure: Effect of amount of stimulus preexposure and coping style. 499
- Shostak, David A., and McIntyre, Curtis W. Stimulus-seeking behavior in three delinquent personality types. 582
- Siegel, Judith M. *See* Sarason, Irwin G.
- Siegelman, Ellen Y. *See* Horowitz, Leonard M.
- Silberfarb, Peter M., Phelps, Patricia J., Hauri, Peter, and Solow, Charles. Effects of intestinal bypass surgery on body concept. 1415
- Silverman, Lloyd H., Martin, April, Ungaro, Roseann, and Mendelsohn, Eric. Effect of subliminal stimulation of symbiotic fantasies on behavior modification treatment of obesity. 432
- Simons, Lynn S., Morton, Teru L., Wade, Terry C., and McSharry, Dennis M. Treatment outcome and follow-up evaluation based on client case records in a mental health center. 246
- Skinner, Harvey A., and Jackson, Douglas N. A model of psychopathology based on an integration of MMPI actuarial systems. 231
- Slaney, Robert B. Therapist and client perceptions of alternative roles for the facilitative conditions. 1146
- Slutsky, Jeffrey M., and Allen, George J. Influence of contextual cues on the efficacy of desensitization and a credible placebo in alleviating public speaking anxiety. 119
- Smith, Charles J. *See* Sutker, Patricia B.
- Smith, Jean L. *See* Goldfried, Marvin R.
- Smith, Harold H., Jr. *See* Heaton, Robert K.
- Smith, Jonathan C. Personality correlates of continuation and outcome in meditation and erect sitting control treatments. 272
- Smith, Ronald E., Johnson, James H., and Sarason, Irwin G. Life change, the sensation seeking motive, and psychological distress. 348
- Snyder, C. R., Handelsman, Mitchell M., and Endelman, Janet R. Can clients provide valuable feedback to clinicians about their personality interpretations? A reply to Greene. 1493
- Sobell, Linda C., and Sobell, Mark B. Validity of self-reports in three populations of alcoholics. 901
- Sobell, Mark B. *See* Sobell, Linda C.
- Solow, Charles. *See* Silberfarb, Peter M.
- Souheaver, Gary T., and Schuldt, W. John. Effects of differences in suggestibility within self- and external control conditions. 1230
- Spinetta, John J. Parental personality factors in child abuse. 1409
- Spinetta, John J., and Maloney, Lorrie J. The child with cancer: Patterns of communication and denial. 1540
- Stake, Jayne E. *See* Walker, Elaine F.
- Stalonas, Peter M., Jr., Johnson, William G., and Christ, Maryann. Behavior modification for obesity: The evaluation of exercise, contingency management, and program adherence. 463
- Stedman, James M., Lawlis, G. Frank, Cortner, Robert H., and Achterberg, Gloria. Relationships between WISC-R factors, Wide-Range Achievement Test scores, and visual-motor maturation in children referred for psychological evaluation. 869
- Steele, Robert E. Steele's reply to Bedrosian and Beck. 1500
- Steinbock, Elizabeth. *See* Lewis, Clifford E.
- Steinert, Yvonne. *See* Campbell, Susan B.
- Stewart, Abigail J. A longitudinal study of coping styles in self-defining and socially defined women. 1079
- Stone, Arthur A., and Neale, John M. Life Event Scales: Psychophysical training and rating dimension effects on event-weighting coefficients. 849
- Stoops, Rick. *See* Sarason, Irwin G.
- Strickland, Bonnie R. Internal-external expectancies and health-related behaviors. 1192
- Sturgis, Ellie T. *See* Forehand, Rex.
- Sturgis, Ellie T., and Adams, Henry E. The right to treatment: Issues in the treatment of homosexuality. 165

- Surwit, Richard S., Shapiro, David, and Good, Michael I. Comparison of cardiovascular biofeedback, neuromuscular biofeedback, and meditation in the treatment of borderline essential hypertension. 252
- Sutker, Patricia B., Archer, Robert P., and Allain, Albert N. Drug abuse patterns, personality characteristics, and relationships with sex, race, and sensation seeking. 1374
- Sutker, Patricia B., Allain, Albert N., Smith, Charles J., and Cohen, Gary H. Addict descriptions of therapeutic community, multimodality, and methadone maintenance treatment clients and staff. 508
- Sutker, Patricia B., Allain, Albert N., and Geyer, Scott. Female criminal violence and differential MMPI characteristics. 1141
- Szapocznik, José, Scopetta, Mercedes Arca, Aranalde, Maria de los Angeles, and Kurtines, William. Cuban value structure: Treatment implications. 961
- Tavormina, Joseph. *See* Gayton, William F.
- Taylor, C. *See* Russell, M. A. H.
- Taylor, Jerome. *See* Terrell, Francis.
- Tellegen, Auke. *See* Butcher, James N.
- Terrell, Francis, Taylor, Jerome, and Terrell, Sandra L. Effect of type of social reinforcement on the intelligence test performance of lower-class black children. 1538
- Tervell, Sandra L. *See* Terrell, Francis.
- Thomson, Richard W. *See* Parker, Jerry C.
- Thurber, Steven, and Torbet, David P. On the word preferences of suicidal versus nonsuicidal college students. 15
- Tinklenberg, Jared R. *See* Renson, Gisele J.
- Torbet, David P. *See* Thurber, Steven.
- Trevithick, Lee, and Hosch, Harmon M. MMPI correlates of drug addiction based on drug of choice. 100
- Tuma, A. Hussain, May, Philip R. A., Yale, Coralee, and Forsythe, Alan B. Therapist experience, general clinical ability, and treatment outcome in schizophrenia. 1120
- Turkewitz, Hillary. *See* O'Leary, K. Daniel.
- Turner, Judith, and McCreary, Charles. Short forms of the MMPI with back pain patients. 354
- Turner, Robert G., and Keyson, Mae. Relationships between client self-perceptions of self-consciousness levels and therapist awareness of these perceptions. 1586
- Ungaro, Roseann. *See* Silverman, Lloyd H.
- Unger, Thomas. *See* Lick, John.
- Vail, Anthony. Factors influencing lower-class black patients remaining in treatment. 341
- Van Hoose, T. A. *See* Penk, W. E.
- Vannicelli, Marsha, Washburn, Stephen, Scheff, Betty-Jane, and Longabaugh, Richard. Comparison of usual and experimental patients in a psychiatric day center. 9
- Vender, Michael. *See* Jeffery, Robert W.
- Vogt, Arthur T. *See* Heaton, Robert K.
- Wade, Terry C. *See* Simons, Lynn S.
- Wagner, Nathaniel N. Is masturbation still sinful? Comments on Bailey's comments. 1507
- Waksman, Steven. Psychometric phenology revisited: Comments on neuropsychological testing. 1486
- Walker, Elaine F., and Stake, Jayne E. Changes in preferences for male and female counselors. 1153
- Waller, Ronald W., and Keeley, Stuart M. Effects of explanation and information feedback on the illusory correlation phenomenon. 342
- Walters, Joyce. Kantorowitz, David A. 586
- Ward, Alan J. Early childhood autism and structural therapy: Outcome after 3 years.
- Washburn, Stephen. *See* Vannicelli, Marsha.
- Watson, Charles G., and Schuld, Donald. Psychosomatic etiological factors in neoplasms: A response to Kellerman. 1524
- Watson, Charles G., and Plemel, Duane. An MMPI scale to separate brain-damaged from functional psychiatric patients in neuropsychiatric settings. 1127
- Weerts, Theodore C., and Lang, Peter J. Psychophysiology of fear imagery: Differences between focal phobia and social performance anxiety. 1157
- Weiss, Joseph. *See* Horowitz, Leonard M.
- Weiss, Robert L. *See* Margolin, Gayla.
- Weiss, Robert L., and Aved, Barbara M. Marital satisfaction and depression as predictors of physical health status. 1379
- Weisz, John R. *See* Glennon, Blair.
- Welling, Martha A. *See* Berzins, Juris I.
- Wells, Karen C. *See* Forehand, Rex.
- Westlake, Robert J. *See* Brownell, Kelly D.
- Wetter, Robert E. *See* Berzins, Juris I.
- Whitman, Douglas. *See* Wilson, Robert S.
- Whyte, Richard. *See* Ottomanelli, Gennaro.
- Wiens, Arthur N. *See* Matarazzo, Joseph D.

Wiggins, Jerry S., and Holzmuller, Ana. Psychological androgyny and interpersonal behavior	40
Wikoff, Richard L. Correlational and factor analysis of the Peabody Individual Achievement Test and the WISC-R.....	322
Wilcox, Stuart A. See Acosta, Frank X.	
Wildman, Robert W., and Wildman, Robert W., II. Validity of the Verbal Scale IQ as a WAIS short form: A reply to Reynolds.....	1537
Wildman, Robert W., II. See Wildman, Robert W.	
Wilkins, Wallace. See Esse, John T.	
Williams, Carolyn L., and Ciminero, Anthony R. Development and validation of a heterosexual skills inventory: The survey of heterosexual interactions for females.....	1547
Williams, Thomas A. See Giannetti, Ronald A.	
Wilson, G. Terence. Ethical and professional issues in sex therapy: Comments on Bailey's "Psychotherapy or massage parlor technology?".....	1510
Wilson, G. Terence. Methodological considerations in treatment outcome research on obesity..	687
Wilson, Peter. See Ottomanelli, Gennaro.	
Wilson, Robert S., Rosenbaum, Gerald, Brown, Gregory, Rourke, Daniel, Whitman, Douglas, and Grisell, James. An index of premorbid intelligence.....	1554
Wing, Rena R. See Jeffery, Robert W.	
Wittmaier, Bruce C. See Post, Amy L.	
Wolfsohn, Lawrence. See Levin, Saul M.	
Wood, Gloria H. See Kurtines, William M.	
Worell, Judith. Sex roles and psychological well-being: Perspectives on methodology ..	777
Wyatt, Richard Jed. See Donnelly, Edward F.	
Yale, Coralee. See Tuma, A. Hussain.	
Yamamoto, Joe. See Acosta, Frank X.	
Young, Veronica. See Kelly, Jeffrey A.	
Yurcheson, R. See Melamed, B. G.	
Zeiss, Robert A. Self-directed treatment for premature ejaculation.....	1234
Ziff, David R. See Newmark, Charles S.	
Zuckerman, Marvin. See Mellstrom, Martin, Jr.	
Zuckerman, Marvin, Eysenck, Sybil, and Eysenck, H. J. Sensation seeking in England and America: Cross-cultural, age, and sex comparisons.....	139
Zukow, Arnold H. See Zukow, Patricia Goldring.	
Zukow, Patricia Goldring, Zukow, Arnold H., and Bentler, P. M. Rating scales for the identification and treatment of hyperkinesis.....	213
Zuroff, David C., and Schwarz, J. Conrad. Effects of transcendental meditation and muscle relaxation on trait anxiety, maladjustment, locus of control, and drug use.....	264
OTHER	
Garfield appointed editor, 1979-1984.....	52, 212, 416
Instructions to authors.....	11, 230, 431, 872
Notice on author alterations.....	1191

CHARLES C THOMAS • PUBLISHER

SEX-RELATED COGNITIVE DIFFERENCES: An Essay on Theory and Evidence by Julia A. Sherman. Biological theories, including purported sex differences in brain lateralization, and related sociocultural determinants are discussed. *cloth-\$16.25, paper-\$12.25*

COUNSELING IN COMMUNICATIVE DISORDERS edited by R. E. Hartbauer. Each chapter focuses on a particular communicative problem, giving a general overview of the disorder, counseling techniques and procedures, and suggestions for evaluation of effectiveness. *\$18.25*

MUSIC THERAPY: An Introduction to Therapy and Special Education Through Music (2nd Ptg.) by Donald E. Michel. A history of music therapy, its basic theoretical rationales, and its applications in therapy, rehabilitation, and special education are presented. *\$11.25*

THE USE OF ALTERNATIVE MODES FOR COMMUNICATION IN PSYCHOTHERAPY: The Computer, The Book, The Telephone, The Television, The Tape Recorder by David Lester. The author discusses each communication mode and its unique psychotherapeutic qualities. *\$12.25*

GROUP COUNSELING AND GROUP PSYCHOTHERAPY WITH REHABILITATION CLIENTS edited by Milton Seligman. Practical applications of group strategies are given particular emphasis along with group methods with vocational or occupational goals. *cloth-\$19.00, paper-\$14.75*

CRISIS INTERVENTION AND HOW IT WORKS by Romaine V. Edwards. The primary focus is on the author's proven method of crisis counseling. *\$9.25*

THEORIES AND METHODS OF GROUP COUNSELING IN THE SCHOOLS (2nd Ed.) edited by George M. Gazda. Topics include group methods for use with the early school child, family group counseling, and Adlerian group consultation. *\$18.25*

A PRIMER ON SCHOOL MENTAL HEALTH CONSULTATION by Morton I. Berkowitz. The basics of the consultation process — how to start school consultation, what goes on once it is in progress, and to what ends it may conclude — are presented. *\$11.50*

MENTAL EXAMINER'S SOURCE BOOK edited by Julian C. Davis and John P. Foreyt. Chapters include the classification of mental retardation, objective and projective personality testing, intelligence testing and the mental status examination. *\$16.75*

THE HAND TEST: A New Projective Test with Special Reference to the Prediction of Overt Aggressive Behavior (5th Ptg.) by Barry Bricklin, Zygmunt A. Piotrowski and Edwin E. Wagner. The text covers daily use of the test in clinical practice, including diagnostic procedures for individual cases. *\$8.25*

THE CARE AND MANAGEMENT OF THE SICK AND INCOMPETENT PHYSICIAN by Robert C. Green, Jr., George J. Carroll and William D. Buxton. The authors detail the management, prevention and identification of problems which impair the ability of a physician to perform adequately. *\$10.00*

BASIC APPROACHES TO GROUP PSYCHOTHERAPY AND GROUP COUNSELING (2nd Ed., 2nd Ptg.) edited by George M. Gazda. Experts in the field summarize such techniques as psychodrama, reality therapy, rational-emotive therapy and gestalt therapy. *\$19.50*

PSYCHIATRIC PROBLEMS IN OPHTHALMOLOGY edited by Jerome T. Pearlman, George L. Adams and Sherwin H. Sloan. Discussions of the eye and the "I," reactions to the loss of sight, and eye symptoms with no organic disease are included along with helpful diagnostic and management procedures. *\$14.50*

HUMAN BEHAVIOR GENETICS compiled and edited by Arnold R. Kaplan. In addition to such topics as basic human genetics, measuring human behavior, and genetic factors in intelligence and personality development, the book shows the need for careful controls and conservative judgement. *\$28.00*

CASE STUDIES OF THE CLINICAL INTERPRETATION OF THE BENDER GESTALT TEST: Illustrations of the Interpretive Process for Graduate Training and Continuing Professional Education by Clifford M. DeCato and Robert J. Wicks. This book establishes the Bender Gestalt Test (BGT) as an effective means for assessing personality dynamics, developmental level and psychopathology. *\$11.25*

TECHNIQUES FOR BEHAVIOR CHANGE: Applications of Adlerian Theory (3rd Ptg.) compiled and edited by Arthur G. Nikelly. Sections include assessment, group and individual therapy, and techniques for dealing with special problems. *\$12.25*

THE EMOTIONALLY DISTURBED CHILD: A Book of Readings (2nd Ptg.) compiled and edited by Larry A. Faas. The text covers man's approach to defining emotional disturbances and offers discussion on the identification and dynamics of emotionally disturbed children. *\$16.75*

Orders with remittance sent, on approval, postpaid

301-327 East Lawrence Avenue • Springfield • Illinois • 62717

Clinical Evaluation of Young Children with the McCarthy Scales

By ALAN S. KAUFMAN Ph.D.
and NADEEN L. KAUFMAN, M.A., Ed.D.

Clinical Evaluation of Young Children with the McCarthy Scales is probably the only book available on the clinical interpretation of the McCarthy test. The authors, who have drawn on years of research and clinical experience with the test, share a psychometric-clinical-learning disabilities orientation towards assessment. In addition, both authors were involved in the development of the scale prior to its publication and have published extensively on the battery during the past five years. Crucial topics such as screening for learning problems, administration tips, and writing, case reports are also featured. Readers and users of this unique work will gain invaluable insight into the McCarthy profile of any children tested.

1977, 320 pp., \$16.50 ISBN: 0-8089-1013-2

The Personal Sphere Model

By RAOUL A. SCHMIEDECK, M.D.

The Personal Sphere Model offers reliable and valid data as an instrument for the assessment of object relationships and for personality research in general. It is a projective test in which interpersonal relationships are presented by drawings. Both a manual for the Personal Sphere Model and an account of its development, this book contains instructions, norms, studies in reliability and validity, examples, and reviews of supporting work.

1978, 240 pp., \$18.50 ISBN: 0-8089-1093-0

Send payment with order and save postage and handling charge.

Prices are subject to change without notice.

Grune & Stratton

A Subsidiary of

Harcourt Brace Jovanovich, Publishers

111 FIFTH AVENUE, NEW YORK, N.Y. 10003
24-28 OVAL ROAD, LONDON NW1 7DX

Contributions to the Psychopathology of Schizophrenia

Edited by BRENDAN A. MAHER

This collection contains reprints of important articles—both empirical reviews and theoretical formulations—on schizophrenia that have appeared in *Progress in Experimental Personality Research*. It provides a comprehensive review of the advances in research on schizophrenia over the past 12 years. All articles except the last are accompanied by postscripts written by the article authors, indicating changes that have occurred in their thinking and new data that bear on their original formulations. Clinical psychologists, psychiatrists, psychoanalysts, and psychopathologists; and abnormal, personality, developmental, cognitive, and experimental psychologists will all find interesting perspectives on their work and the history of its development in these seminal articles.

CONTENTS: P. H. Venables, Input Disfunction in Schizophrenia. Postscript. L. J. Chapman et al., A Theory of Verbal Behavior in Schizophrenia. Postscript. I. I. Gottesman, Contributions of Twin Studies to Perspectives in Schizophrenia. Twin Studies and Schizophrenia a Decade Later. W. E. Broen, Jr., and L. H. Storms, A Theory of Response Interference in Schizophrenia. Response Disorganization and Narrowed Observation. J. M. Neale and R. L. Cromwell, Attention and Schizophrenia. Postscript. P. Magaro, Theories of the Schizophrenic Performance Deficit.

1977, 400 pp., \$11.25 ISBN: 0-12-465250-6

Progress in Experimental Personality Research

VOLUME 6

Edited by BRENDAN A. MAHER

FROM THE PREVIEWS:

"... unusually excellent scholarly and closely reasoned review... highly recommended."

—AMERICAN JOURNAL OF PSYCHIATRY

"... bound to be arresting to many readers, and further demonstrates that experimental psychology is 'not so much a subject as a way of life' in the reduction of scientific uncertainty."

—Vernon Hamilton in NATURE

This serial publication is designed to synthesize current significant developments in the study of personality for psychologists, psychopathologists, psychiatrists, sociologists, educators, and other students of behavioral science. Contributions range from original works in important technical areas to summaries of data and integrated reviews of present knowledge.

1978, 368 pp., \$23.00 ISBN: 0-12-541408-0

Send payment with order and save postage and handling charge.

Prices are subject to change without notice.

ACADEMIC PRESS, INC.

A Subsidiary of

Harcourt Brace Jovanovich, Publishers

111 FIFTH AVENUE, NEW YORK, N.Y. 10003
24-28 OVAL ROAD, LONDON NW1 7DX

AHSEN'S EPT

Description: Ahsen's Eidetic Parents Test (EPT) elicits eidetic images of parents, in the mind of the subject, in 30 situations. Each response is a triadic experiential unit — an image(I), a somatic response(S), and a meaning(M). A precise test for localizing personality structures, somatic and emotional patterns and psychological themes in the form of images which do not change unless transformed through therapeutic dynamics.

"A methodological advance"

— *American Journal of Psychiatry*

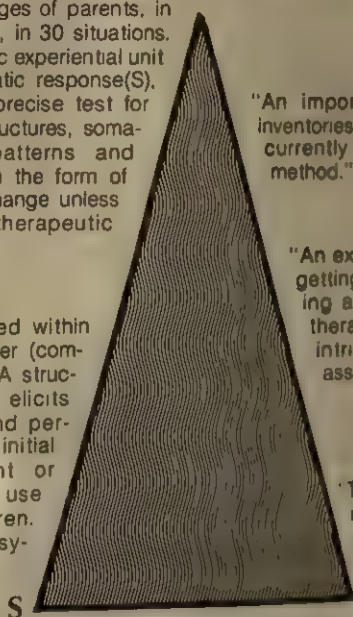
"An important addition to the fear inventories and reinforcement schedules currently in use...a rapid and effective method."

— *Behavior Therapy*

"An exciting and ingenious way of getting at conflict areas and providing a rich source of material for therapeutic sessions. A new and intriguing device for personality assessment"

— *Contemporary Psychology*

Uses: Easily administered within one hour (brief) or longer (comprehensive), if desired. A structured interview which elicits valid developmental and personality data, also an initial diagnostic instrument or therapy process, for use with adults and children. Full instructions for psychological report writing.



"Therapeutic as well as diagnostic and fact finding" — *The International Journal of Clinical and Experimental Hypnosis*

EIDETIC PARENTS TEST

- ▲ CLINICAL DIAGNOSIS
- ▲ PERSONALITY ASSESSMENT
- ▲ DEVELOPMENTAL FUNCTIONING
- ▲ SPECIFIC THERAPY IMAGES

for

➤ **BRANDON HOUSE, Inc.** For orders, write PO Box 240, Dept D Bronx N.Y. 10471

Yes, send me Akhter Ahseen's **EIDETIC PARENTS TEST AND ANALYSIS**. I am enclosing \$17.50 plus 50¢ postage and handling. (N.Y.S. residents, please add 8% tax.)

Name _____

Address _____

City _____

State _____

Zip _____

☐ Send bill (institutions only)

Methodology in Clinical Research

Special issue of the Journal of Consulting and Clinical Psychology August 1978

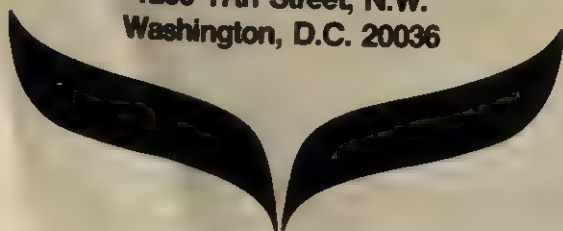
Major methodological aspects of clinical topic areas are described and discussed. Topic areas include smoking and addiction, marital and child treatment, sex roles, obesity, and others.

Contributors include the editor, Brendan A. Maher, Alan E. Kazdin, Richard M. McFall, Peter E. Nathan, David Lansky, and Judith Worell to name a few.

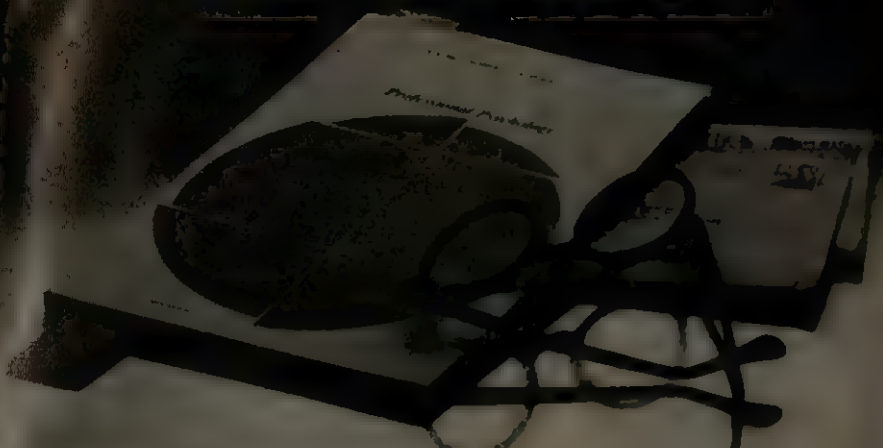
Copies of the special issue are available at \$6 each, prepaid. Discounts on bulk orders are also available. To order single copies or for more information on bulk orders write:



Subscription Department
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036



NOW: INFORMATION ABOUT THE LAW AND PSYCHOLOGY.



It's here:

Law and Professional Psychology. This special August 1978 issue of **Professional Psychology** brings you thirteen original articles full of fresh thought and constructive recommendations on the issues that vitally affect your day-to-day practice. You'll find precautions you may take to protect yourself as a professional and a plan for improving investigations of malpractice claims. Expert witness testimony, jury selection, civil commitment, confidentiality and privilege, minor's consent to treatment, and the use of psychological devices are critically examined.

Send your order today. And soon you'll receive this special issue.

Single copies of the issue are available for \$5 each. All orders \$25 or less must be prepaid. Make and send checks payable to:

American Psychological
Association
Order Department
1200 17th Street, N.W.
Washington, D.C. 20036

Enclosed is \$ _____ for _____ copies of **Professional Psychology's** special issue—Law and Professional Psychology—at \$5.00 each.

NAME _____

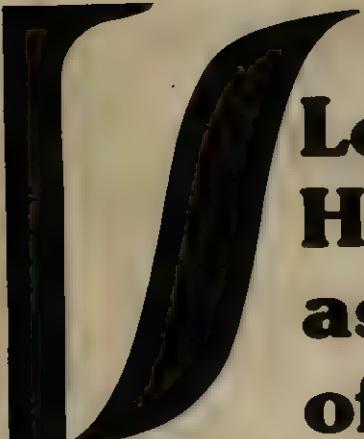
ADDRESS _____

CITY _____

STATE _____

ZIP CODE _____

pp-18



Learned Helplessness as a Model of Depression

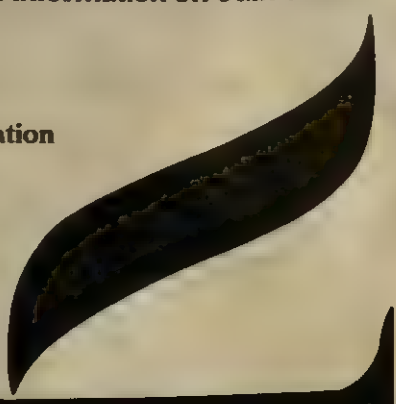
**Special issue of the
Journal of Abnormal Psychology,
February 1978.**

The issue serves as a forum to investigate the validity and adequacy of the learned helplessness model of depression in humans and includes both critical and supportive theoretical and empirical articles with comments.

L. Rowell Huesmann is guest editor. Richard A. Depue, Martin E. P. Seligman, Charles G. Costello, Robert C. Smolen, and Alexander M. Buchwald are among the contributors.

Copies of the special issue are available at \$6 each, prepaid. Discounts on bulk orders are also available. To order single copies or for more information on bulk orders write:

**Subscription Department
American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036**



Have the life of your time!

A lot of your colleagues do!

For the past 13 years, many members of the APA have taken advantage of their eligibility to purchase life insurance at special group-rate savings. And each year they've found that they get more out of life — because the low cost term rates they pay have been further reduced by not less than 20% in premium credits. A reduction that is unheard of in today's inflationary economy. You can join your colleagues by applying for life

insurance coverage in amounts ranging from \$10,000 to \$100,000. If you already have life insurance, this is an inexpensive way to add to the coverage you now have. And you get the additional benefit of accidental death and dismemberment coverage. Also, there's an optional life coverage available for your spouse and children.

**So don't wait. Fill out the coupon below and send it in now, because —
Now is the time to have the life of your time.**

Mr. Harold Belcher, Administrator
APA Insurance Plans
Liberty Life Assurance Company of Boston
175 Berkeley Street
Boston, Massachusetts 02117

Please send me more information about APA group-rate life insurance.

NAME _____

ADDRESS _____

CITY _____

STATE _____ ZIP _____

Do you have other insurance through APA with Liberty? ☐ Yes ☐ No

(If you're a resident of either Ohio or Texas, information about special low-rate individual life policies will be sent. These states do not permit group life contracts for associations like the APA.)

LIBERTY LIFE

**ASSURANCE COMPANY
OF BOSTON**
HOME OFFICE BOSTON



79

Need an individual test which *quickly* provides a stable and reliable estimate of intelligence in 4 or 5 minutes per form? Has three forms? Does not depend on verbal production or subjective scoring? Can be used with the severely physically handicapped (even paralyzed), if they can signal yes-no? Handles two-year-olds and superior adults within the same short series of items and the same format? Has been painstakingly standardized? Builds rapport with intrinsically interesting and simple procedures based on multiple drawings?

Ammons and Ammons

QUICK TEST (QT)

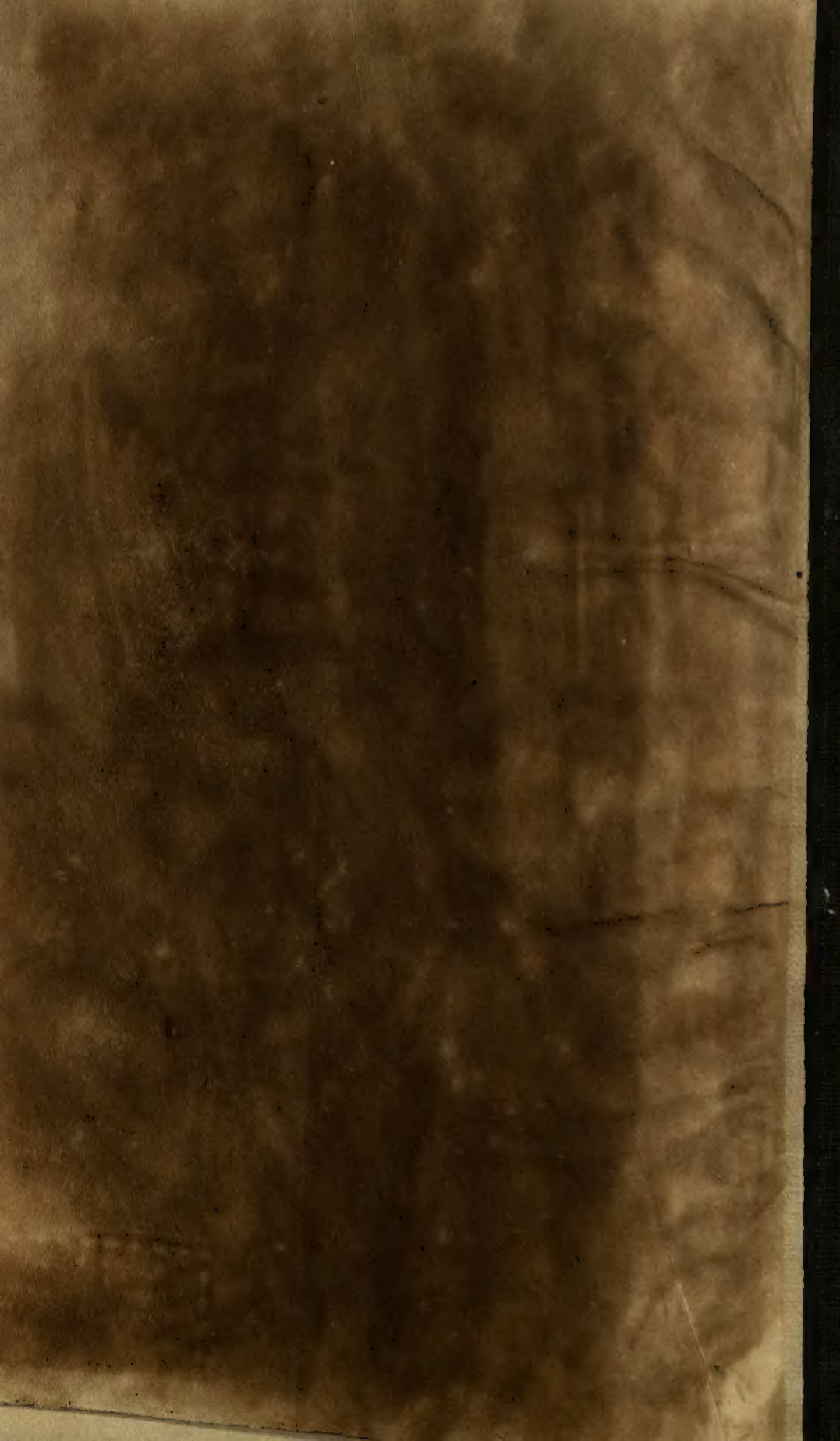
Order Tester's Set including manual, card materials, keys, and 100 answer sheets (\$16.00 complete).

For further information, write to

**PSYCHOLOGICAL TEST
SPECIALISTS**

Box 9229

Missoula, Montana 59807



MADE BY THE DISABLED
REHABILITATION - INDIA
CALCUTTA-700 053

